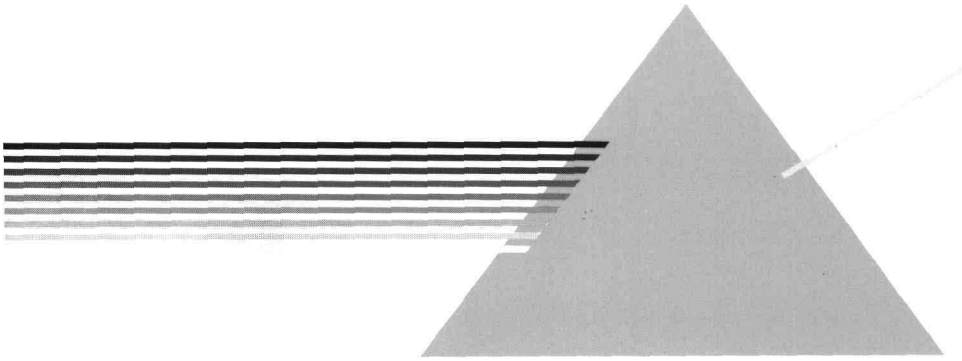
MURIEL AMAR

LES FONDEMENTS THÉORIQUES DE L'INDEXATION

UNE APPROCHE LINGUISTIQUE



ADBS
ÉDITIONS

Muriel Amar

Les Fondements théoriques de l'indexation

Une approche linguistique

ADBS Editions

Cet ouvrage est le texte d'une thèse de doctorat en Sciences de l'information et de la communication préparée sous la direction de Michel Le Guern et soutenue le 11 décembre 1997 à l'Université Lumière Lyon 2 devant un jury composé de : Danièle Dubois, directeur de recherche au CNRS (InaLF, UPR 9017), présidente du jury ; Richard Bouché, professeur à l'ENSSIB (École nationale supérieure des sciences de l'information et des bibliothèques), rapporteur ; Jean-Paul Metzger, professeur à l'Université Jean Moulin Lyon 3, rapporteur ; Michel Le Guern, professeur à l'Université Lumière Lyon 2, directeur de thèse ; et Jean-Marc Proust, conservateur en chef à l'ENSSIB. Cette thèse a obtenu la mention très honorable avec félicitations du jury à l'unanimité.

Conservateur de bibliothèques, Muriel Amar exerce actuellement à Médiadix, centre régional de formation aux carrières des bibliothèques (Médiadix - Pôle Métiers du Livre de l'Université Paris X-Nanterre, 11 avenue Pozzo di Borgo, F-92210 Saint-Cloud, courriel Muriel.Amar@u-paris10.fr), où elle est notamment chargée des formations en nouvelles technologies. Sa formation antérieure en documentation (DESS Information et documentation de l'Institut d'études politiques de Paris) lui a permis de développer une réflexion sur l'indexation qui embrasse les deux problématiques documentaire et bibliothéconomique. Elle poursuit actuellement ses recherches sur l'indexation et la catégorisation au sein de l'équipe LCPE (Langage, cognition, pratiques et ergonomie / Institut national de la langue française, CNRS UPR 9017), animée par Danièle Dubois.

ISBN 2-84365-042-9 - ISSN 1159-7666

La voie qui peut s'énoncer

N'est pas la voie pour toujours

Le nom qui peut la nommer

N'est pas le nom pour toujours.

Lao Tseu, *Tao te King*, 1.

SOMMAIRE

PRÉFACE

Il y a au moins cinq cents ans que l'on indexe, et pourtant il aura fallu attendre le milieu du XX[e] siècle pour voir apparaître les mots *indexer* et *indexation*, que les dictionnaires datent de 1948. On a commencé par des ouvrages, et de là on est passé à des collections. Au début du XVI[e] siècle, ce n'est pas encore le mot *index* qui désigne le résultat de l'opération ; c'est *tabula*. Il s'agit bien de rompre la linéarité des documents traités, d'en donner une projection tabulaire, qui permette d'y tracer un chemin autre que celui que les auteurs avaient choisi. C'est ainsi, en partant de l'index, que la plupart des lecteurs se promenaient dans Pline l'Ancien, dans les *Essais* de Montaigne ou dans les *Commentaires hiéroglyphiques* de Pierius Valerianus.

En tête du tome V de ses *Diversités* (1610), Jean-Pierre Camus, l'évêque de Belley, ami de saint François de Sales, dit son hostilité à la pratique de l'indexation et au mode de lecture qu'elle induit. Il demande au lecteur de ne pas considérer comme une imperfection le fait que son livre soit « sans Indice des mémorables » : « C'est une erreur populaire, qui n'infecte que les faibles cerveaux, qui appellent cela l'âme du livre, et c'est l'instrument de leur stupidité. Ces gens peuvent être appelés *Doctores tabularii*, lesquels *sapiunt tantum per Indices*. Les enquerrez-vous de ce qu'ils savent ? Ils vous demandent un livre pour le montrer, et aussitôt à la Table pour trouver ce qu'ils cherchent, les habiles appellent cela le pont aux ânes. » Les quatre premiers volumes des *Diversités* étaient munis d'index, d'ailleurs fort bien faits, ce qui n'arrête pas les protestations de Jean-Pierre Camus : « Les tables des tomes précédents de l'auteur, faites par je ne sais qui, et à son insu, lui déplaisent, sachant qu'il faut retrancher tant que l'on peut ce qui fomente la paresse, paresse mère de l'ignorance. » Les volumes suivants comportent des index. Le fait que la protestation de Jean-Pierre Camus soit restée vaine, même auprès de ses propres éditeurs, montre que l'indexation répond à un véritable besoin, dès que l'imprimerie a multiplié les documents : on ne peut pas tout lire, de tous les livres, même en n'étant pas paresseux. À la nécessité empirique de trouver de l'information répond la pratique de l'indexation, qui restera empirique pendant plus de quatre siècles.

On a donc pu indexer, génération après génération, sans même éprouver la nécessité de nommer cette pratique, et *a fortiori* de la théoriser. Cela ne présentait pas d'inconvénient majeur dès lors que l'indexation était la tâche d'un homme seul : la qualité de l'indexation était fonction de la qualité de l'indexeur, et nombreuses sont les tables qui ne manquent ni de rigueur ni de cohérence. La limite de la masse de documents à indexer n'était limitée que par la puissance de travail de celui qui

s'en chargeait. On demandait à Du Cange, dont les glossaires sont sans doute la plus vaste entreprise d'indexation de l'époque classique, comment il avait pu mener à bien une telle tâche. C'était en y travaillant depuis l'âge de dix-huit ans, tous les jours, douze heures par jour, à une seule exception près : le jour de son mariage, il n'avait travaillé que huit heures.

La réduction des horaires et l'accroissement des collections condamnaient l'indexation à devenir une tâche collective, en posant de redoutables problèmes de cohérence. Il a donc fallu normaliser. Mais un savoir-faire empirique ne se laisse pas facilement normaliser, et les normes sans fondements sont les pires des ornières. Au mieux, tant qu'il ne s'agit que de coordonner des tâches qui restent purement humaines, on peut arriver à un semblant de cohérence, même si l'essentiel du consensus reste dans l'implicite. Mais, le jour où le recours aux moyens automatiques a imposé une rationalisation et une explicitation totale des procédures, force a été de constater qu'une approche empirique, même sous sa forme normalisée, ne suffisait plus. Il fallait donner à la vieille pratique de l'indexation, enfin nommée, des fondements théoriques.

Les premiers travaux sur l'indexation automatique ont cherché à faire simuler par la machine les procédures manuelles ; au mieux, on faisait moins bien. On continuait à penser dans le cadre d'un système documentaire qui répondait à une requête de l'utilisateur par une liste de références. Les progrès technologiques de ces dernières années ont tout bouleversé : la possibilité de transférer le document sur des supports informatiques a comme conséquence que l'utilisateur ne se contente plus de la référence ; il lui faut le texte lui-même. L'indexation manuelle était nécessairement partielle, alors que les moyens actuels permettent de viser à l'exhaustivité dans le traitement de l'information contenue dans les documents. Les choix du documentaliste intervenaient dans la détermination des éléments à retenir pour l'indexation ; ils portent aujourd'hui sur la sélection des documents à indexer.

Mais, de l'indexation manuelle à l'indexation automatique, il reste une continuité : la nature du descripteur reste fondamentalement la même, et les structures cognitives de l'esprit humain n'ont pas changé. L'évolution de l'outil rend toutefois nécessaire de fonder la pratique sur une épistémologie explicitée. Le livre de Muriel Amar répond à cette nécessité.

Au centre de la construction, l'indexation est située comme une pratique discursive. Ce ne sont pas les mots qui sont importants, mais les choses que ces mots désignent. Et les mots de la langue ne sont mis en relation avec les choses que par le discours. En outre, le fait de considérer l'indexation comme une pratique discursive est pleinement justifié par le fait que l'univers de discours du documentaliste ne coïncide ni avec ceux des auteurs ni avec ceux des utilisateurs.

Quant au rapprochement de l'indexation et de la vulgarisation scientifique, il ouvre une perspective nouvelle et féconde. Il permet de prendre conscience, sans sombrer dans le désespoir, d'une évidence que la plupart des praticiens et des théoriciens de l'indexation ont refusé de voir : le postulat qui veut que l'univers réel soit le même pour tous les acteurs de la chaîne documentaire, depuis les auteurs des textes sources jusqu'aux utilisateurs, est manifestement faux. Muriel Amar contourne la difficulté par le recours à la théorie des mondes possibles de Kripke, qui permet d'articuler les divers univers de discours.

L'analyse théorique est suivie d'une ouverture sur des propositions pratiques, qui s'écartent des usages habituels. Il n'est jamais confortable de soumettre à un examen

critique la *doxa* officielle de tout un milieu professionnel, même avec un point de vue extérieur, mais l'appartenance de l'analyste à la profession dont il risque de déranger ainsi les habitudes exige un véritable courage intellectuel – pas seulement intellectuel, d'ailleurs. Les compétences de Muriel Amar en épistémologie et en linguistique garantissent la pertinence de son analyse, et son appartenance au corps des conservateurs amplifie la portée de sa démarche. L'efficacité est ici à la mesure du risque accepté.

Michel Le Guern
Professeur à l'Université Lumière Lyon 2

AVANT-PROPOS

L'indexation telle qu'elle a été pratiquée par des générations entières de bibliothécaires et de documentalistes ne serait-elle qu'un simple savoir-faire ?

Certes les entreprises d'indexation, de la plus modeste à la plus aboutie, ont toujours cherché à faire système, à se protéger des subjectivités douteuses, ont toujours redouté la sémantique d'une époque et tenté de trouver, au-delà, les termes aptes à se confronter à l'éternité, à porter en eux, malgré et contre le temps, une forme d'universel et d'immanence. De telles entreprises ont donc toujours été empreintes d'une recherche de rigueur intellectuelle, d'un souci de cohérence, animées par une volonté opiniâtre de constituer des clés d'accès aux savoirs, à la pensée et à l'expression humaines.

Mais, nous dit Muriel Amar dans ce livre fort et exigeant, nous sommes-nous jamais réellement donné les moyens de nos ambitions ? Avons-nous réellement interrogé nos pratiques ? Avons-nous réellement passé nos méthodes au crible de l'analyse scientifique ? Nous qui nous penchons si souvent, justement pour les indexer, sur des travaux scientifiques, avons-vous vraiment cherché à constituer un savoir fondé sur une démarche scientifique ou avons-nous laissé la place à une approximation, à terme fautive ?

Par la seule puissance de son analyse, Muriel Amar donne le sentiment que l'indexation est peu à peu devenue une pratique, on pourrait presque dire une coutume, tellement intégrée dans nos activités fondatrices (mais routinières) qu'elle en devient une évidence – apparente – qui n'est que fort peu interrogée.

Muriel Amar appelle à ce qu'on pourrait appeler une refondation de la légitimité de l'indexation, balayant au passage les arguties de ceux qui, toujours fascinés et dominés par les modes, voudraient faire accroire que la révolution électronique rendrait inutile une telle tâche. Bien au contraire ! Elle ne fait qu'en accroître l'urgence...

Une telle refondation passe par une analyse épistémologique. Muriel Amar défend une approche non instrumentale de l'indexation, fondée sur la linguistique, et clairement différenciée de la recherche documentaire. Le descripteur possède alors une forme d'autonomie par rapport au texte qu'il entend nommer, et doit être apte à *« conjoindre la stabilité de la signification avec l'instabilité de la désignation ».*

Dans un des chapitres les plus séduisants de sa recherche, s'appuyant tout à tour sur Paul Ricœur et Michel Foucault, Muriel Amar cherche les voies d'une indexation conçue comme un aller et retour entre décontextualisation et recontextualisation des documents, passant par une reconnaissance de l'*« épaisseur discursive »* des textes, et capable de construire des descripteurs aptes à refléter tout à la fois la singularité et la pluralité d'un texte, son unicité tout comme son appartenance à tous les textes.

La tentation est grande alors, de répondre à l'invitation de Muriel Amar. Si le descripteur idéal se donne comme ce qui permet *« de circuler dans un espace documentaire conçu a priori comme homogène »,* n'est-il pas alors à lui seul, et d'une manière finalement cohérente, la métaphore même de la bibliothèque ? Ce n'est pas le moindre mérite du livre de Muriel Amar que de remettre l'indexation, dont l'importance est alors pleinement assumée, au cœur des enjeux d'un exercice professionnel.

Martine Poulain
Directrice de Médiadix, Université Paris X

INTRODUCTION

Cette recherche se donne pour objectif de fonder, d'un point de vue théorique, une pratique professionnelle exercée principalement dans les bibliothèques et les centres de documentation : l'indexation, habituellement définie par les praticiens comme « l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document[1] ».

Une étude des fondements théoriques de l'indexation suppose la constitution, à partir d'un objet empirique (la pratique d'indexation), d'un objet scientifique[2].

Cet objectif soulève un certain nombre de problèmes méthodologiques :

(i) concernant l'objet empirique : en quoi les pratiques professionnelles d'indexation peuvent-elles constituer un objet empirique ? Peuvent-elles s'appréhender de façon unifiée ?

(ii) concernant l'objet scientifique : quel point de vue scientifique, quelle science peuvent permettre de construire l'indexation en tant qu'objet scientifique ?

(iii) concernant le type de relation qu'entretiennent objet empirique et objet scientifique : s'il est possible de déterminer des fondements théoriques à l'indexation, qu'est-ce qu'une telle étude peut permettre d'apprendre de l'objet empirique dont elle prétend rendre compte ?

Avant de proposer un cadre de réponse à ces trois questions, nous présenterons succinctement les enjeux d'une étude des fondements théoriques de l'indexation. Nous indiquerons en fin d'introduction le plan suivi dans cette recherche.

---

[1] Norme AFNOR (Association française de normalisation) Z 47-102 (1978), p. 225.

[2] L'opposition proposée ici entre « objet empirique » et « objet scientifique » s'appuie sur l'opposition classique entre *Techné* et *Epistémé* ; les deux notions se distinguent du point de vue de la nature de leur objet : « instable » dans le cas de la *Techné*, « stable » dans le cas de l'*Epistémé*, voir Granger 1993, chapitre II.

I - Enjeux d'une étude des fondements théoriques de l'indexation

Si cette recherche entreprend, en marge des traités théoriques sur l'indexation[1], de repenser la question sous l'angle des fondements théoriques, c'est essentiellement sous « la pression technologique » que connaissent les domaines professionnels de l'information et de la communication. Il s'avère en effet que les descriptions classiques de la pratique d'indexation manuelle dont disposent les professionnels ne leur permettent pas toujours de se situer dans les débats, nouveaux ou moins nouveaux, qui portent par exemple sur l'indexation automatique[2] ou sur l'usage de moteurs d'indexation à l'œuvre sur le réseau Internet[3]. Se multiplie parallèlement une littérature consacrée, elle, à l'automatisation des procédures d'indexation[4], sans que ne soit toujours rendu de façon nette le rapport que l'on peut établir entre l'indexation documentaire, telle que les professionnels la réalisent, et l'indexation automatisée, telle que des systèmes peuvent la produire. D'une certaine façon, on est passé de l'indexation manuelle à l'indexation automatique sans que le niveau d'une appréhension formelle des questions ait été réellement établi. Sur ce point, la problématique de l'indexation ne dépare pas celle des autres objets qui constituent le champ des sciences de l'information et de la communication[5].

Cette carence d'approche théorique de l'indexation n'a pas, semble-t-il, posé de problèmes majeurs jusqu'à ce que :

- d'une part, le nombre croissant de systèmes d'indexation automatisés et la somme totale des coûts engagés n'aient alerté, notamment les pouvoirs publics, sur la nécessité de procéder à une évaluation de ces systèmes ;

- d'autre part, les procédures mises en œuvre sur le réseau Internet imposent de nouveaux modes de traitement des documents qui ne doivent rien à l'indexation telle qu'elle se laisse couramment décrire.

Dans les deux cas, l'absence d'approches formelles de l'indexation constitue un véritable obstacle à l'appréhension comme à la discussion des enjeux mettant en cause l'indexation. Cependant, ce n'est pas toujours la voie d'une entreprise de constitution théorique de l'indexation qui a été retenue ; c'est au contraire une mise à distance de l'indexation elle-même qui semble se dessiner.

A - Place de l'indexation dans les méthodes d'évaluation

L'évaluation de l'indexation a suivi de près la naissance de l'indexation elle-même en tant que pratique professionnelle reconnue. Des outils et méthodes d'évaluation

---

[1] Proposés par exemple par Lancaster [1991] et Fugmann [1993].

[2] Voir, par exemple, sur ce point, Le Moal 1997, p. 380-384. Pendant longtemps, les « logiciels documentaires » (les logiciels dédiés au traitement des documents) sont restés entre les seules mains des informaticiens ; ils passent ensuite dans celles des linguistes-informaticiens avec le développement de l'« ingénierie de l'information ».

[3] Voir, par exemple, sur ce point, Michel 1997, p. 361-363 : « Internet est aussi une vision nouvelle du libre parcours dans l'information. Il pose donc des problèmes à l'industrie de l'information et remet en cause des pratiques développées depuis une vingtaine d'années ».

[4] Pour une synthèse, Sidhom [thèse en cours].

[5] Sur ce point, Le Coadic [1997, p. 516] fait remarquer que « dans le domaine de l'information, la connaissance technique a souvent précédé la connaissance scientifique ».

ont été constitués concomitamment aux outils et méthodes de l'indexation ; les premiers, inspirés des seconds, reposent pour l'essentiel sur une mesure des résultats que l'on peut obtenir à partir de requêtes documentaires[1]. La transposition de ces outils et méthodes à l'évaluation des systèmes d'indexation automatisés s'est avérée problématique[2] et a conduit à la mise en œuvre de programmes de recherche portant sur les méthodes d'évaluation elles-mêmes. Ainsi, pour la période récente, peut-on citer, entre autres, les programmes de recherche suivants :

- sur le plan européen, la Direction Générale XIII (*Language and Research Engineering*) de la CEE a lancé le programme *Eagles*, dont l'un des axes consiste à développer des méthodes d'évaluation pour les produits et services de traitement linguistique de l'information ;

- dans le cadre d'actions multilatérales francophones, l'AUPELF-UREF[3] a engagé une action de recherche concertée (*Amaryllis*) dont le but est de « permettre à la fois à la recherche de progresser et au domaine de se doter d'instruments de mesure rendant possible une comparaison des différentes approches[4] » ;

- en France, une réflexion a été lancée par le CNRS dans le cadre du programme *GRACE*. Parallèlement, le ministère de l'Enseignement supérieur et de la Recherche constitue depuis peu son propre programme de recherche dans le domaine[5].

Les méthodes d'évaluation de l'indexation existantes, sans doute valables lorsqu'elles sont utilisées par les professionnels dans le cadre singulier de leur pratique, ne révèlent pas la même pertinence dès lors qu'il s'agit d'évaluer des systèmes reposant sur des modèles, implicites le plus souvent, de l'indexation : on a bien du mal, dans ces cas, à trouver l'« aune de référence » qui permette de les discuter.

En l'absence d'approche formelle de l'indexation qui permettrait d'évaluer les systèmes d'indexation automatisés sous l'angle d'une évaluation de modèles, la problématique de l'évaluation de l'indexation tend à céder le pas à une évaluation de la capacité des systèmes automatisés à permettre une « bonne » recherche d'information : ce n'est plus l'indexation comme telle qui est évaluée, c'est plutôt un ensemble de procédures diverses censées répondre au même objectif qu'elle[6]. Par conséquent, la question de la « consistance » de l'indexation se pose : l'indexation telle que les praticiens l'exercent n'est-elle qu'une des techniques possibles parmi d'autres ? Dans ce cas, l'indexation peut-elle constituer un objet

---

[1] Sur ce point, voir, dans le glossaire, les entrées « taux de rappel » et « taux de précision ».
[2] Pour une synthèse, on peut se reporter à Sparck-Jones (éd.) 1981. Par exemple, p. 1 : « There is no very good reason to suppose that the conventional methods are best, even in principle, let alone practice » ; p. 3 : « It is arguable that our current understanding of information processing is like of sixteenth century herbolists : it embodies some observation and insight, but lacks detailed analysis and supporting theory ».
[3] AUPELF : Association des universités partiellement ou entièrement de langue française ; UREF : Université des réseaux d'expression française.
[4] AUPELF-UREF 1994, annexe, [p. 1].
[5] Chaudiron 1994, p. 100-104.
[6] Sur ce point, le programme américain TREC (*Text REtrieval Conference*) est exemplaire. On trouvera une présentation des objectifs de ce programme et des résultats auxquels il permet d'aboutir dans Lespinasse 1997.

d'étude ? N'est-ce pas plutôt l'ensemble des procédures utilisées en recherche d'information qui doit alors être analysé ?

B - Place de l'indexation dans le réseau Internet

La constitution du réseau Internet s'est accompagnée de la création d'un ensemble d'outils spécifiques[1], dont certains jouent le rôle de l'indexation documentaire. Ainsi de ceux que l'on appelle les « moteurs de recherche[2] » : ce sont « des bases de données constituées automatiquement grâce à des logiciels appelés robots qui scrutent à intervalles réguliers les serveurs déclarés sur Internet. [...] Ils indexent mot à mot les documents localisés permettant ainsi des interrogations par sujets.[3] » Ces robots qui « indexent » ne procèdent aucunement à une « représentation des concepts » contenus dans les documents, pour reprendre le texte de la norme. Le type d'indexation mis en œuvre ne ressemble en rien à l'indexation que les professionnels pratiquent[4]. Comment interpréter cet état de fait ? L'indexation telle que la pratique professionnelle la définit, la décrit, l'exerce n'est-elle qu'une technique conjoncturelle, liée à un état de la technologie aujourd'hui dépassé, sans être une opération fondamentalement liée au processus de transfert d'information ? L'indexation documentaire est-elle une opération nécessaire ou une technique simplement utile ? Là encore, quel objet le chercheur doit-il retenir pour son étude : l'indexation ? les divers procédés permettant la recherche d'information ?

On aurait tort, nous semble-t-il, d'évacuer trop rapidement les problématiques spécifiques de l'indexation en les diluant dans celles de la recherche documentaire ou dans celles des nouvelles technologies de l'information, c'est-à-dire sans avoir préalablement essayé de formaliser ce qui constitue en propre l'indexation. Alors que « la pression technologique » actuelle tend à laisser les objets se fondre et se confondre (indexation et recherche documentaire, notamment), cette étude entend donner les moyens de « reconnaître » l'indexation sous les aspects différents que peuvent lui donner l'histoire d'une profession comme celle des techniques avec lesquelles elle évolue. Ces moyens, de nature théorique, doivent permettre d'analyser l'indexation en toute généralité, mais aussi de pouvoir capter son évolution, et, pourquoi pas, de la prévoir.

Certes, les pratiques d'indexation telles qu'elles se laissent voir et décrire ne semblent guère sujettes à des généralisations de cet ordre ; il nous semble cependant possible de constituer l'indexation comme un objet scientifique présentant certaines caractéristiques de stabilité.

II - L'indexation, un objet empirique

Qui cherche à étudier l'indexation dispose d'un ensemble d'observatoires de nature différente.

---

[1] Dont le World Wide Web et les techniques associées, entre autres le format HTML (Hyper Text Markup Language).
[2] Comme, par exemple, Alta Vista, Excite ou Lycos.
[3] Lardy 1994, p. 6.
[4] Pour le détail, on peut se reporter à Le Crosnier 1996.

On peut étudier l'indexation sur les lieux professionnels où elle s'exerce et auprès des indexeurs : c'est alors la façon dont les indexeurs indexent qui est analysée. On peut étudier l'indexation telle que les systèmes informatiques la simulent ou cherchent à en simuler les résultats : on s'intéresse alors soit aux procédures techniques (capacité de traitement, temps de réalisation, etc.) soit aux formes de modélisation, implicites et explicites, qui sont à l'œuvre (modélisations de nature mathématique, linguistique, cognitive, etc.). On peut enfin étudier l'indexation telle qu'elle est décrite dans la littérature (normative, didactique, scientifique, etc.).

Une fois déterminés ces différents « lieux » d'inscription de l'indexation (professionnel, technique, discursif), on peut spécifier l'angle d'approche retenu : le processus de l'indexation (les opérations qui la composent), son résultat (souvent appelé descripteur), son objet (le document), ses outils (les langages documentaires), ses supports (on parlera alors de l'indexation de texte, de l'indexation d'image, de l'indexation de carte, de phonogramme, de vidéogramme, etc.).

Le processus lui-même de l'indexation peut être considéré d'au moins deux façons : comme une opération « englobée » ou comme une opération « englobante ». L'indexation peut se concevoir comme une des formes de réalisation, possible parmi d'autres, de l'« analyse de contenu » (à côté de la classification, du résumé et de la synthèse documentaire, etc.). Elle peut également être appréhendée elle-même comme une analyse de contenu, qui se spécifie par le type d'outil qu'elle utilise (classification, langage documentaire, « langage naturel », représentation « conceptuelle », etc.).

Sur la base de cet aperçu, non exhaustif, des aspects de l'indexation, se dégage la diversité des approches possibles. Comme toute pratique sociale, professionnelle, l'indexation ne peut constituer en tant que telle un objet d'analyse. Il est clair qu'une seule approche ne saurait rendre compte de l'ensemble des problématiques de l'indexation. En cela, la pratique de l'indexation constitue typiquement l'objet d'une interdiscipline comme les sciences de l'information et de la communication[1], qui permettent, par le biais de différents types de théorie, de « découper » un aspect du « réel » de l'indexation, et donc de se doter d'un objet empirique observable, sur la base duquel pourra se construire ultimement, par la convergence des approches, un objet scientifique.

III - L'indexation, un objet de quelle science ?

La pratique professionnelle de l'indexation se laisse décrire par le biais de normes, traités, manuels, du point de vue particulier de la pratique elle-même, dans le cadre d'un référentiel proprement documentaire qui maintient l'indexation dans la

[1] Les sciences de l'information et de la communication ont pu être définies comme une « interdiscipline centrée sur l'étude des processus de l'information et de la communication relevant d'actions organisées, finalisées, prenant appui ou non sur des techniques et participant d'actions sociales et culturelles », Comité National d'Évaluation 1993, p. 123. Sur ce point, on peut aussi consulter Têtu [1997, p. 513-516] et Le Coadic [1997, p. 516-523].

complexité de ses manifestations, soumise à une diversité de facteurs de nature hétérogène (institutionnel, technique, historique, etc.).

L'intérêt d'étudier l'indexation dans le cadre des sciences de l'information et de la communication tient au fait que l'ensemble des disciplines[1] par lesquelles elles se constituent comme science permet de disposer d'un ensemble de points de vue théoriques différents et distincts : chacun de ces points de vue propose un référentiel spécifique permettant d'analyser, à un niveau qui lui est propre, l'un des multiples aspects en jeu dans une pratique professionnelle.

Cependant, travailler l'indexation dans le cadre d'une interdiscipline ne présente pas que des avantages[2]. De nombreuses difficultés sont à prendre en considération : comment s'articulent les différents points de vue sur un objet si le « réel » qu'ils permettent de découper n'est pas exactement le même ? Si, du point de vue de la discipline considérée, on peut espérer « tout » voir du phénomène observé, comment savoir ce que ce point de vue permet de faire voir de la globalité de la pratique retenue pour étude ? Pour un objet empirique donné, y a-t-il des approches disciplinaires plus légitimes que d'autres ?

Sur ce point, toute recherche entreprise dans le cadre des sciences de l'information et de la communication doit, nous semble-t-il, contribuer à proposer des réponses à ces questions. Nous essaierons, quant à nous, de prendre en compte cet aspect de la problématique des sciences de l'information et de la communication.

Parmi les différentes disciplines qui constituent les sciences de l'information et de la communication, se trouve la linguistique qui, si elle a été depuis longtemps sollicitée pour l'étude des faits d'indexation, n'a pas toujours été invoquée pour conduire une étude théorique de l'indexation. Comme le note Janik [1985] dans son bilan des rapports entre linguistique et sciences de l'information, la littérature abondante, qui, à la fin des années 60 et aux débuts des années 70, a porté sur les aspects linguistiques des processus documentaires, a surtout privilégié, en fait, le point de vue de l'indexation automatique. Ainsi des ouvrages de Bély et *al.* [1970], Coyaud et *al.* [1972], Cros, Gardin et *al.* [1964], pour les plus connus.

Pour des raisons qui seront développées dans le premier chapitre de cette recherche, nous retiendrons, pour conduire notre étude des fondements théoriques de l'indexation, la linguistique comme discipline de référence. Précisons d'ores et déjà que le choix de cette approche a été déterminé par le travail mené depuis plus de quinze ans par Michel Le Guern et les membres de l'équipe SYDO[3] : les travaux entrepris permettent de disposer d'acquis à partir desquels peuvent se formuler, aujourd'hui, les fondements de l'indexation du point de vue de la théorie linguistique. À bien des égards, cette recherche ne constitue qu'une synthèse,

---

[1] Les sciences de l'information et de la communication se constituent à partir de plusieurs champs disciplinaires, notamment : économie/droit, anthropologie/sociologie, psychologie, linguistique, logique/statistiques/mathématique, histoire/épistémologie/philosophie. Voir par exemple Le Coadic 1994 pour un essai de clarification.

[2] Les sciences de l'information et de la communication constituent un champ de recherche récent (elles existent institutionnellement depuis 1975) dont l'unité épistémologique reste encore en discussion. L'ensemble de ses concepts n'est pas, pour le moment, entièrement établi, Comité National d'Évaluation 1993, p. 87.

[3] L'équipe SYDO (pour SYstèmes DOcumentaires) était composée, à ses débuts, de Alain Berrendonner, Richard Bouché, Sylvie Lainé, Michel Le Guern, Jean-Paul Metzger, Jacques Rouault.

menée du seul point de vue de l'indexation, des études menées par l'équipe SYDO. Nous rappelons donc dans ses grandes lignes le programme de recherche qui a été suivi[1], les acquis qui nous serviront de base de travail et la contribution que voudrait apporter cette étude.

A - Programme de recherche de l'équipe SYDO

Si l'équipe SYDO a travaillé dans le cadre de l'indexation automatique[2], pour laquelle elle a construit un analyseur morpho-syntaxique[3], d'emblée s'est imposée la nécessité de disposer d'un modèle de description formelle de l'unité que l'analyseur visait à extraire des textes[4]. Le descripteur a en effet fait l'objet d'une formalisation détaillée[5], qui s'appuie sur la mise en valeur de ses propriétés spécifiques, référentielle et discursive, justifiant son approche en tant que « syntagme nominal[6] ». Le cadre d'analyse retenu pour rendre compte du fonctionnement particulier du descripteur est double : à la fois linguistique[7] et logique[8]. Dans ce cadre, des études ont pu être menées dont les perspectives d'automatisation reposaient, de façon constante, sur une approche des faits de langue[9].

B - Les acquis

Les études menées dans le cadre logico-sémantique établi par Michel Le Guern permettent aujourd'hui de disposer d'acquis théoriques, qui ne sont pas sans jeter de nouveaux éclairages sur les pratiques d'indexation classiques :

(i) la notion de langage documentaire a pu être mise à distance sur la base d'une étude du rôle des « mots » en indexation : la dimension lexicale propre au

---

[1] La présentation, ici succincte, des travaux réalisés par l'équipe SYDO sera reprise en détail dans la suite de cette recherche.

[2] Le Guern 1994, p. 75 : « Conçu en vue de l'indexation automatique, l'analyseur morpho-syntaxique élaboré par l'équipe SYDO a eu comme premier objectif l'extraction de tous les syntagmes nominaux présents dans le texte à indexer, ces syntagmes nominaux étant amenés à jouer le rôle des descripteurs dans le système d'information. La première tâche a consisté à établir un système de règles qui permette de reconnaître les syntagmes nominaux dans des documents à indexer, le syntagme nominal étant défini, dans une perspective où se croisent la grammaire et la logique, comme la plus petite unité de discours susceptible de servir de base à une relation référentielle autonome ».

[3] La grammaire de l'analyseur a été établie par Berrendonner 1983 et Metzger 1988.

[4] Le Guern 1991a, p. 22 : « Ce dont je suis sûr [...] c'est que vouloir appliquer la linguistique aux traitements automatiques sans se préoccuper de modèles, c'est courir à l'échec, même si le bricolage habile d'un bon informaticien un peu teinté de linguistique peut faire illusion un certain temps. [...] L'informatisation des systèmes documentaires impose la nécessité d'une réflexion théorique sur les opérations qui en constituent les composantes. [...] Le passage de l'indexation manuelle à l'indexation automatique ne modifie pas la nature des descripteurs, mais il oblige à ne plus se contenter d'une approche intuitive et empirique. On peut indexer à la main sans savoir exactement ce qu'est un descripteur ; en revanche, on ne peut pas mettre en place un système d'indexation automatique sans une réflexion préalable sur les descripteurs, et sans une certaine formalisation ».

[5] Le Guern 1984 notamment.

[6] Sur ce point, voir Bouché 1989.

[7] Voir Le Guern 1997, p. 375-379.

[8] Voir Metzger 1997, p. 385-390.

[9] Sans pouvoir être exhaustive, on peut citer, par exemple, sur le traitement des anaphores dans une perspective documentaire, Vidalenc-Sabourin [1989], sur le traitement des conjonctions de coordination, Larouk [1994].

langage documentaire ne correspond pas à la dimension discursive en jeu dans l'indexation ;

(ii) l'indexation se laisse décrire sous la forme d'une extraction d'unités de discours : les notions de « représentation de concepts » et de « traduction de concepts » sont alors à revisiter ;

(iii) la recherche documentaire se laisse, elle aussi, redéfinir[1] : elle a pour finalité non plus l'appariement de « mots », mais plutôt la détermination d'objets particuliers que sont les objets de discours.

Par ailleurs, des rapprochements inédits et fructueux ont pu être opérés entre documentation et terminologie[2], engageant là aussi la recherche dans des voies de nature à spécifier le descripteur sous l'angle des propriétés qu'il partage avec le terme de la terminologie.

Cet important travail de mise au jour des propriétés du descripteur fournit des pistes d'exploration pertinentes pour parcourir le vaste champ des travaux linguistiques à la recherche d'éléments pour fonder la pratique d'indexation du point de vue de la théorie linguistique. En ce sens, les deux dimensions, référentielle et discursive, du descripteur, mises en valeur par Michel Le Guern et les membres de l'équipe SYDO, ont permis de guider notre investigation dans le champ linguistique.

C - Notre contribution

Compte tenu de notre parcours antérieur, nous avons privilégié, dans cette étude des fondements théoriques de l'indexation, le versant linguistique des hypothèses proposées par l'équipe SYDO[3]. Nous avons donc exploré des modes de représentation linguistique de la référence en général et de la référence discursive en particulier, en privilégiant un ensemble de travaux qui s'inscrit de façon plus ou moins lâche dans le programme de recherche proposé par Milner [1989][4].

À partir de ce référentiel linguistique, nous avons repris les hypothèses émises par l'équipe SYDO pour les reformuler dans le cadre des problématiques de l'indexation.

L'essentiel du travail mené sous la direction de Michel Le Guern a porté sur le descripteur vu sous l'angle de la recherche d'information. Nous nous sommes, quant à nous, plus particulièrement attachée au descripteur vu sous l'angle de l'indexation proprement dite, en élargissant la problématique au processus de l'indexation lui-même. Cet angle d'analyse permet de proposer des fondements théoriques concernant en propre l'indexation.

---

[1] Sur ce point, on peut suivre Kuramoto [1995 et thèse en cours].
[2] Le Guern 1989, Mustafa-Elhadi 1989.
[3] Y compris les lectures linguistiques des modèles issus de la logique.
[4] Ce cadre propose une reformulation du programme de recherche proposé par Chomsky. Toutefois, toutes les études linguistiques sur lesquelles nous nous appuierons dans cette recherche ne relèvent ni du même cadre ni du cadre précisément spécifié par Milner. Cependant, elles ne contredisent pas l'option retenue par ce dernier.

IV - Rapport entre objet empirique et objet scientifique

Comme nous le préciserons dans le premier chapitre de cette étude, notre problématique – l'étude des fondements de l'indexation du point de vue d'une théorie linguistique – nous conduit à privilégier, parmi la multiplicité empirique par laquelle peut se capter l'indexation, les discours sur l'indexation[1]. En ce sens, cette recherche porte non pas sur la façon dont les indexeurs ou les systèmes automatisés indexent, mais sur les arrière-plans théoriques sur lesquels reposent de telles pratiques, manuelles ou machinales, d'indexation.

L'étude de cet arrière-plan théorique, tel qu'il se manifeste dans les discours sur l'indexation, permet de constituer, en partie, l'indexation comme objet scientifique : c'est sur la base de reformulation des modèles implicites de la langue que l'on peut spécifier les propriétés linguistiques en jeu dans l'indexation. Sur ce point, nous rejoignons les propositions de Gardin sur le rôle que l'on peut faire tenir aux théories dans l'étude de pratiques non formelles : « À quoi bon prendre la peine de formaliser ou de programmer la collecte et la structuration des données par des voies dont rien ne garantit *a priori* qu'elles se révéleront plus fécondes ou plus « intéressantes », pour l'archéologue ou l'historien, que les voies dites traditionnelles ? N'est-il pas plus raisonnable d'inverser la stratégie, c'est-à-dire de choisir d'abord un certain nombre de théories que la communauté savante ou du moins une partie d'entre elle, tient pour intéressantes ou fécondes, puis d'en donner une version formelle dans l'espoir que l'appareil cognitif ainsi dégagé bénéficiera par construction de ces mêmes qualités, pour d'autres emplois ?[2] »

Notre étude des fondements théoriques de l'indexation présente donc cette particularité d'approcher l'objet empirique « indexation » par le biais d'interrogations issues de problématiques linguistiques : pourquoi est-ce des noms, des unités nominales, qui sont depuis toujours et partout utilisés en indexation, et pas, par exemple, des unités verbales ? Quelle différence y a-t-il entre un descripteur « nom propre » et un descripteur « nom commun », entre un descripteur composé d'un mot et un descripteur composé de plusieurs mots ? Pourquoi les pratiques d'indexation recourent-elles invariablement à la langue, alors même que les discours sur l'indexation ne cessent d'en pointer « l'imperfection », « l'ambiguïté » ? Comment l'indexation appréhende-t-elle la spécificité sémiotique des objets qu'elle manipule, que ce soit les textes qu'elle sélectionne ou les univers de discours qu'elle permet de traverser ?

Cet ensemble de questions, qui ne se pose que dans le cadre d'une approche linguistique des faits d'indexation, permet, par touches, de faire émerger les propriétés de langue sur lesquelles reposent les pratiques professionnelles d'indexation. À ce pouvoir explicatif d'une approche linguistique de l'indexation s'adjoint un pouvoir de nature plus prédictive : on peut déterminer des « manières d'indexer » de diverses natures, dont certaines sont à même de tirer harmonieusement profit des progrès technologiques sans s'y laisser dissoudre.

Le matériau utilisé dans cette recherche des fondements théoriques de l'indexation sera donc principalement constitué d'un ensemble de discours sur la pratique

---

[1] Cet angle d'étude de l'indexation est couramment retenu ; voir, par exemple, Dubois 1995 ou Van Holland 1995.
[2] Gardin 1991, p. 24.

d'indexation[1], auquel s'ajoute une dimension expérimentale. Nous avons en effet réalisé une enquête auprès de dix organismes documentaires[2] dans le but :

(i) d'analyser le mode d'exploration des sources en indexation : comment l'indexeur construit-il son objet d'indexation, le document ?

(ii) d'étudier le rapport entre le type de document construit et le type de formule d'indexation établi : quelle est l'incidence de la « mise en document » en indexation ?

Les interprétations auxquelles cette enquête peut donner lieu reposent sur les hypothèses que nous permet de formuler une approche linguistique de l'indexation : celles-ci seront spécifiées en cours d'étude ; de même, les conclusions auxquelles on peut aboutir seront ponctuellement rapportées au fil du texte, en fonction des aspects de l'indexation étudiés.

V - Plan de la recherche

Cette recherche procède en trois temps.

- Le premier chapitre est consacré à la formulation de notre problématique. Il propose un cadre qui permet de traiter la question des fondements théoriques de l'indexation. Pour cela, il précise l'objet étudié et la méthode d'analyse retenue, en répondant notamment aux trois questions suivantes :

(i) comment l'indexation peut-elle constituer un objet d'étude spécifique ?

(ii) en quoi une approche en termes de « fondements théoriques » paraît-elle plus adaptée à l'objet étudié qu'une approche en termes de « théorie » proprement dite ?

(iii) pourquoi retenir, parmi l'ensemble des approches possibles, le point de vue de la théorie linguistique ?

- Les deux chapitres suivants sont regroupés dans une première partie intitulée « problèmes théoriques de l'indexation ». Deux problèmes théoriques y sont abordés : la question du lexique en indexation fait l'objet du chapitre II, celle de la référence l'objet du chapitre III. Sur ces deux questions, on examine respectivement le point de vue des professionnels et le point de vue des linguistes, en s'interrogeant sur les zones de distorsion entre descriptions et sur les zones de désaccord entre modes d'appréhension.

- Les chapitres IV et V constituent la seconde et dernière partie de cette étude, intitulée « contribution aux fondements théoriques de l'indexation ». On y propose une reformulation de l'indexation qui repose sur un modèle explicite

---

[1] Le texte-pivot de cet ensemble de discours est le discours normatif (norme Z 47-100).
[2] Présentée en annexe 1 ; les principaux résultats sont rapportés dans les annexes 2 et 3.

de la langue. Les propriétés linguistiques pertinentes en indexation sont pensées dans le cadre d'un modèle qui permet de les « utiliser » à des fins professionnelles. L'indexation est, dans cette seconde partie, appréhendée sous ses deux aspects de processus et de résultat : le chapitre IV propose de considérer le processus de l'indexation comme un mode d'organisation spécifique des documents, un niveau de « discours » particulier. Le chapitre V reprend la problématique du descripteur sous l'angle d'une approche discursive de l'indexation.

L'articulation de cette recherche est plus précisément présentée à la fin du chapitre I, la formulation de notre problématique permettant de spécifier la logique d'exposition retenue.

Par ailleurs, chacune des deux parties fait l'objet d'une introduction et d'une conclusion spécifiques.

Les termes suivis d'une étoile (*) renvoient au glossaire pages 329 et suivantes.

CHAPITRE I
EXPOSÉ DE LA PROBLÉMATIQUE

Ce chapitre a pour objectif de présenter un cadre de recherche dans lequel la question des fondements théoriques de l'indexation puisse être posée.

En effet, le sujet de notre recherche n'est pas si évidemment pertinent : il suppose une approche de l'objet d'étude – l'indexation – et un type d'analyse – permettant de capter les fondements théoriques – qui coïncident peu avec la représentation de l'indexation telle qu'on la trouve traditionnellement exprimée dans la littérature.

Notre recherche repose en effet sur deux présupposés :

- d'une part, l'indexation peut faire l'objet d'une étude spécifique, c'est-à-dire constituer, en elle-même, un objet de recherche (I) ;

- d'autre part, une approche de l'indexation qui se veut théorique ne peut donner corps à une théorie* de l'indexation ; elle peut en revanche tenter d'établir les fondements théoriques de l'indexation (II).

I - Définir l'objet d'étude : approches de l'indexation

S'il paraît particulièrement essentiel de définir, dans cette recherche, un cadre pour penser notre objet d'étude, c'est que l'indexation en tant que telle ne peut constituer un objet de recherche dans la perspective de toutes les approches.

Après avoir situé la problématique de l'indexation comme objet d'étude, nous examinerons les propositions de la littérature classique ; nous en dégagerons les limites, à partir desquelles nous proposerons un autre mode d'appréhension de l'indexation.

I.1 - L'indexation, un objet d'étude ?

Que l'indexation puisse constituer un objet d'étude ne va pas de soi ; la difficulté peut être ainsi exprimée :

- sur un plan méthodologique, un objet d'étude doit pouvoir être considéré de façon autonome : il faut pouvoir « isoler » un phénomène[1] ;

- sur un plan épistémologique, un objet étudié dans la perspective des sciences de l'information et de la communication est nécessairement un objet appréhendé dans sa finalité[2].

La question qui se pose est alors la suivante : quelle place accorder à la finalité de l'indexation ? Est-elle définitoire du processus de l'indexation, ou plutôt, comment est-elle définitoire de l'indexation ?

De ce point de vue, on peut opposer :

- une approche « instrumentale » : la définition de l'indexation est établie sur la base de la question « à quoi sert-elle ? ». Le point de vue sur l'objet est externe. Dans cette approche, l'indexation peut difficilement être pensée comme un objet d'étude dans la mesure où elle n'est pas « isolable » ;

- une approche « procédurale » : la définition de l'indexation est établie sous l'angle de la question « comment fonctionne-t-elle ? ». Le point de vue sur l'objet est interne. La finalité de l'indexation se laisse alors déduire de son mode de fonctionnement : c'est l'indexation elle-même qui porte son « mode d'emploi », la possibilité de son usage, sa finalité. À ce titre, elle peut constituer un objet d'étude spécifique. C'est l'hypothèse que nous défendrons.

I.2 - Approches classiques de l'indexation

Les approches classiques* proposent une définition « instrumentale » de l'indexation qui peut prendre la forme suivante : l'indexation « a pour but de faciliter l'accès au contenu d'un document ou d'un ensemble de documents à partir d'un sujet ou d'une combinaison de sujets (ou de tout autre type d'entrée utile à la recherche)[3] ». Traditionnellement, l'indexation est définie de cette façon : comme un outil au service d'une fonction, la recherche documentaire*.

---

[1] Granger [1993, p. 72] parle de « réduction des phénomènes aux objets de science ».

[2] Rappelons que les sciences de l'information et de la communication s'attachent à « l'étude des processus de l'information et de la communication relevant d'actions organisées, *finalisées* », Comité National d'Évaluation 1993, p. 123 (c'est nous qui soulignons). La notion d'« action finalisée » peut être entendue comme « action née d'un besoin social », Granger 1993, p. 33.

[3] Pomart et Sutter 1997, p. 284. Ce type de définition se trouve aussi dans le discours normatif (voir ci-après) mais aussi dans les discours didactiques [par exemple, Chaumier 1996, p. 18 : « Description du contenu du document à l'aide de mots-clés (ou indices de classification) pour faciliter la mémorisation du contenu de ce document *pour une recherche ultérieure* »] et dans les discours techniques [par exemple, Menon 1988, p. 146 : « Identification et enregistrement des unités d'information minimales pertinentes *pour*

Le problème vient de ce que la finalité de l'indexation, lorsqu'elle est formulée par la notion de recherche documentaire, introduit une « circularité » d'analyse : la recherche documentaire est, en effet, elle-même définie par rapport à l'indexation.

La circularité de cette approche est particulièrement visible dans le texte de la norme :

- l'une des finalités de l'indexation est de permettre la recherche d'information : « la finalité de l'indexation est de permettre une recherche efficace des informations contenues dans un fonds de documents et d'indiquer rapidement, sous forme concise, la teneur d'un document » ;

- la recherche d'information est contrainte par l'indexation : « l'indexation conduit à l'enregistrement des concepts contenus dans un document, sous une forme organisée et facilement accessible, c'est-à-dire à la confection d'outils de recherche documentaire. [...] La recherche des informations enregistrées [...] s'opérera à partir de ces outils de recherche documentaire.[1] »

On comprend aisément pourquoi l'indexation est ainsi définie par les professionnels : l'indexation représente, pour eux, un moyen de remplir leur mission qui est, entre autres, de permettre de retrouver des documents. L'indexation intéresse d'abord le professionnel sous l'angle de ce à quoi elle sert.

Cependant, il n'est pas sûr que cette approche « instrumentale » de l'indexation permette de l'étudier, ou du moins, de la constituer comme objet d'étude. En effet, l'approche de la finalité de l'indexation en termes de recherche documentaire s'avère problématique :

- le modèle de la recherche documentaire qui sert implicitement de référence à ce type de définitions se révèle être un modèle partiel (A) ;

- la recherche documentaire apparaît en outre comme une finalité « seconde » de l'indexation : c'est la notion plus large de « service à rendre » qui constitue le fond des approches classiques de l'indexation (B) ;

- la question de la recherche documentaire s'avère n'être enfin qu'une « inversion » de celle de l'indexation : la symétrie qui se dessine entre indexation et recherche documentaires conduit à rendre les objets indistincts (C).

A - Un modèle partiel de la recherche documentaire : le modèle de l'« Information Retrieval *»*

L'approche classique de l'indexation qui met en avant, comme l'indique la norme, la « confection des outils de recherche » (c'est-à-dire la confection de langages documentaires*), adopte implicitement un modèle de la recherche documentaire fondé sur la technique de l'appariement entre les mots d'une requête et les mots issus de l'analyse d'un document. Comme l'a notamment mis en valeur Kolmayer[2],

---

apporter des réponses aux requêtes présentées au système d'information »]. C'est nous qui soulignons.

[1] Norme AFNOR Z 47-102 (1978), p. 225.

[2] Kolmayer 1995, p. 27-33.

cette technique de l'appariement relève du modèle de l'*Information Retrieval*, modèle dominant dans les années 1970, qui voient se développer les systèmes d'interrogation informatisés. Or depuis, et notamment grâce aux analyses cognitives des situations de recherche d'information[1], ce modèle de la recherche documentaire s'est révélé inadéquat : il est notamment apparu que le besoin d'information, ne restant pas constant au cours d'une session de recherche, ne pouvait se trouver exprimé une fois pour toutes dans une requête. C'est ainsi que, depuis le milieu des années 1980 surtout, le problème de la recherche documentaire a pu être reformulé : « S'agissant de l'accès à l'information, les dernières années ont vu la problématique de l'appariement d'une requête à un ensemble de documents se positionner progressivement à l'intérieur d'un cadre plus vaste, celui de la satisfaction du besoin d'information de l'utilisateur. La recherche d'information est devenue alors un problème de recherche cognitive.[2] »

Il peut paraître excessif (et un peu rapide) de considérer la recherche documentaire comme relevant exclusivement d'une approche cognitive ; cependant, il semble tout à fait nécessaire de prendre en compte les résultats des analyses cognitives des situations de recherche : si la pratique d'indexation ne se justifie que pour rendre possible la recherche documentaire, comment pourrait-elle ignorer ces nouvelles représentations des situations d'interrogation ? L'indexation telle qu'elle se pratique reste-t-elle compatible avec ce que l'on connaît désormais des modes de recherche ? Ne se trouve-t-elle pas remise en cause et amenée à se redéfinir ? C'est le sens de l'interrogation que formule Sylvie Lainé-Cruzel dans ces termes : « L'approche dominante concernant l'indexation postule que le sens est contenu dans le document et que, pour aboutir à une représentation du sens, il faut analyser aussi finement que possible la structure apparente du texte, c'est-à-dire identifier et caractériser (par des outils linguistiques et statistiques) la forme prise par le texte, comme étant le reflet exact du sens que l'auteur a donné au texte. Parallèlement émergent de nouveaux modèles, encore au stade théorique, qui s'appuient sur l'idée que le sens est construit par le lecteur, qu'il est différent d'un lecteur à l'autre, et qu'une bonne représentation du contenu passe par la prise en compte de certaines caractéristiques du futur utilisateur. Les approches sont-elles complémentaires ? antinomiques ?[3] »

Si l'on peut admettre que la norme pour l'indexation des documents[4], élaborée à des fins professionnelles en 1978, ait pu prendre implicitement pour cadre le modèle dominant de la recherche documentaire à l'époque (celui de l'*Information Retrieval*), on peut être plus surpris de relever que ce type de représentation – partiel – des situations de recherche documentaire n'ait pas conduit depuis à une redéfinition de l'indexation.

Cependant, à y regarder de près, il apparaît que la finalité de l'indexation exprimée en termes de recherche documentaire n'est, en fait, que « seconde ». C'est implicitement dans le cadre d'un objectif de communication plus large que l'indexation est en réalité définie. C'est cet objectif qui contraint le modèle de la recherche documentaire à être réduit, dans les approches classiques, à celui de l'*Information Retrieval*.

---

[1] Pour la bibliographie, on peut se reporter à Kolmayer 1997.
[2] Dachelet 1990, p. 24.
[3] Lainé-Cruzel 1994, p. 143.
[4] Norme AFNOR Z 47-102 (1978).

B - Hypothèse implicite sur les objectifs de la communication : l'« hypothèse de service »

La formule « hypothèse de service » a été proposée par Escarpit[1] pour exprimer une des limites du modèle mécaniste de l'information.

En montrant que l'approche classique de l'indexation met en œuvre cette hypothèse de service, on pourra mettre au jour le modèle mécaniste dans lequel implicitement elle s'inscrit.

Rappelons que le modèle mécaniste de l'information, qui repose sur une formule mathématique que Shannon a développée en 1948, propose une représentation linéaire de la communication, dans laquelle le transfert de l'information s'effectue de la source au destinataire de la façon suivante :

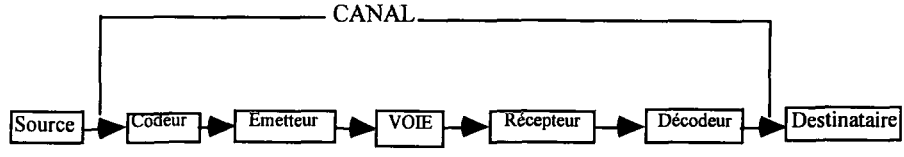


Figure 1 – Le schéma linéaire de la communication
Escarpit 1991, p. 27

Dans cette représentation, seul le destinataire est concerné par la signification des messages transmis. Celui qui véhicule les messages doit être, comme l'est un employé des Postes[2], insensible au contenu : le messager est simplement chargé de rendre un service, c'est-à-dire de « transmettre le plus rapidement, le plus économiquement et surtout le plus fidèlement possible au destinataire l'information contenue dans les messages de la source.[3] »

Cette notion de service à rendre devient une hypothèse – et une hypothèse douteuse, dit Escarpit[4] – quand elle est sollicitée par les tenants du discours mécaniste pour ne pas avoir à s'exprimer ni sur la source de l'information (quelle est-elle ? elle-même de l'information ?), ni sur le message, ou plutôt sur la forme du message (quelle est sa nature ? symbolique ?). Autrement dit, et la critique d'Escarpit porte sur ce point, « ce que voudrait faire accepter ce que nous nommons l'hypothèse de service : [c'est que] l'employé des Postes qui achemine le télégramme n'est pas responsable de ce qui se passe dans l'esprit de l'expéditeur ou du destinataire[5] ». L'hypothèse de service revient donc à poser que le transfert de l'information ne participe lui-même ni à la fabrication de l'information ni à l'interprétation que le destinataire peut en faire[6].

---

[1] Escarpit 1991, p. 44-46.
[2] Escarpit [*Ibid.*, p. 30-31] reprend ici l'image qu'a proposée Roubine dans son *Introduction à la théorie de la communication* [1970].
[3] *Ibid.*, p. 45.
[4] *Ibid.*, p. 44-45.
[5] *Ibid.*, p. 46.
[6] *Id.*

En examinant les définitions classiques de l'indexation, on relève la présence de l'hypothèse de service mise en lumière par Escarpit :

- l'indexation se donne comme une réponse technique à un problème technique de transmission d'information, sans qu'elle ait à se préoccuper du contenu de ce qu'elle transmet. L'enjeu de l'indexation a en effet pu être ainsi exprimé : « comment dans l'expression d'un sujet préserver la similitude du sens à travers la variété et les incertitudes des langues naturelles ?[1] » ; or ni la question du « sujet » ni celle du « sens » ne sont véritablement posées ;

- l'approche classique ne s'exprime pas sur la source de la transmission (l'indexeur) pas plus que sur la forme du message (les langages documentaires) : ni l'indexeur ni les langages documentaires n'apparaissent comme des éléments susceptibles d'influer sur la réception de l'information. L'indexation se donne comme une opération neutre et transparente de transfert d'information.

Il apparaît donc que les approches classiques de l'indexation reposent implicitement sur une représentation particulière de la recherche documentaire dans la mesure où celle-ci permet de maintenir valide le modèle présupposé de la transmission transparente de l'information.

Or, si le modèle mécaniste a été reconnu essentiel pour rendre compte du fonctionnement du « canal » dans la chaîne de communication, il a été parallèlement jugé insuffisant pour considérer et la source et la forme du message[2]. Ces faiblesses ont conduit les chercheurs en sciences de l'information et de la communication à reprendre et à modifier ce modèle inspiré de Shannon ; cependant, toutes les approches de l'indexation ne semblent pas avoir intégré ces critiques. Est-il envisageable de considérer que cette non-prise en compte soit liée au caractère « vicié[3] » de l'hypothèse de service qui, si elle met bien en avant la question du « service » à rendre, permet aussi de ne pas avoir à en dire plus ?

L'adoption implicite du modèle mécaniste s'avère en outre fâcheuse en ce qu'elle tend à rendre indistinctes indexation et recherche documentaires : les deux opérations se donnent en effet comme les deux faces d'un même processus[4].

C - Hypothèse implicite de la symétrie

Les approches classiques de l'indexation admettent implicitement ce que nous appelons l'hypothèse de la symétrie, hypothèse qui permet de poser la notion d'« indexation des requêtes » et de la traiter au même titre et dans la même forme que l'indexation des documents[5]. Cette hypothèse conduit à rendre indistinctes indexation et recherche documentaires. Si l'indexation et la recherche constituent les deux faces d'un même processus, l'objet d'étude, dans le cadre des approches classiques, ne saurait être ni l'indexation ni la recherche documentaires mais le

---

[1] Maniez 1993, p. 254.
[2] Escarpit 1991, p. 45.
[3] Escarpit avance la notion de « duperie » pour qualifier les démarches qui établissent une « frontière entre le canal et ce qui vient avant et après lui ». *Ibid.*, p. 46.
[4] Par exemple, Hodge [1992, p. 11] : « Indexing and information retrieval can be viewed as two sides of the same coin ».
[5] Voir la norme AFNOR Z 47-102 (1978), p. 225 : l'« indexation de la question, opération analogue à l'indexation des fonds ».

« processus » censé rendre compte de ces deux pratiques. Or, si ce processus existe, on ne le trouve ni nommé ni décrit dans la littérature du domaine[1].

Il est clair que cette hypothèse de symétrie ne tient que dans le cadre des modèles précédemment explicités :

- la technique de l'appariement impose en effet un traitement des questions dans les mêmes termes que ceux utilisés pour l'indexation des documents[2] ;

- l'hypothèse de service dans laquelle s'inscrit l'indexation suppose une mise à distance du « contenu » véhiculé, mise à distance qui autorise un traitement indifférencié des documents et des questions, les deux étant appréhendés sans distinction en termes de message[3].

L'hypothèse de la symétrie, qui donne la possibilité de parler d'indexation de requêtes, apparaît donc plus comme une conséquence de l'adoption implicite du modèle mécaniste que comme une volonté positive de traiter sur un même plan requêtes et documents. Preuve en est que cette conséquence semble bien embarrassante aux yeux mêmes de ceux qui pratiquent l'indexation des requêtes.

On peut en effet relever, dans la littérature, un silence, évocateur nous semble-t-il, sur la question de l'indexation des requêtes. Si elle est toujours citée dans les manuels, traités théoriques et normes, c'est toujours secondairement, de façon succincte et par analogie avec l'indexation des documents[4]. En outre, les quelques auteurs qui la traitent explicitement ne manquent pas de relever l'absence de symétrie absolue entre les deux processus :

- ainsi Chaumier [1996], par exemple, met-il en avant, dans l'indexation des questions, l'importance de la « maïeutique », ceci pour répondre à la critique des professionnels se plaignant que « l'utilisateur ne sait pas ce qu'il veut » quand il formule une requête. L'indexation des questions nécessite, précise-t-il, un dialogue susceptible de permettre à l'utilisateur de trouver les « bons » mots ;

- pour les auteurs qui travaillent dans le cadre d'une automatisation de l'indexation, la symétrie de traitement est encore moins évidente. Ainsi Menon[5], s'il assimile sous le terme de « texte » documents et requêtes, établit-

---

[1] Du moins dans la partie de la littérature du domaine que nous avons explorée, voir la bibliographie.

[2] AFNOR Z 47-102 (1978), p. 225 : « L'utilisation d'un langage documentaire pour ces deux opérations permet d'obtenir une coïncidence exacte du vocabulaire d'indexation des documents avec celui des questions auxquelles ces documents apportent une réponse ».

[3] Cette indifférence au contenu, propre au modèle mécaniste, est clairement exprimée par Roubine, cité *in* Escarpit 1991, p. 31 : « La signification des messages n'est pas prise en considération. [...] Il n'y a aucun inconvénient à identifier texte et message ».

[4] Nous n'avons pas trouvé d'ouvrages ou d'articles traitant exclusivement l'indexation des requêtes. Par contre, dans les ouvrages ou articles abordant explicitement l'indexation des documents, on trouve toujours quelques phrases sur l'indexation des requêtes. L'analyse de la norme AFNOR est, de ce point de vue, éclairante : sous le titre « Principes pour l'indexation des documents », il est bien question, dans la quasi-totalité de la norme, de l'indexation des documents, à l'exception d'une phrase (p. 225) qui, presque incidemment, indique qu'un traitement analogue pourra être fait sur les questions des utilisateurs.

[5] Menon 1988, n. 7, p. 152.

il une différence de procédure : dans le cas du document, l'indexation porte sur des textes longs et doit être sélective ; dans le cas de la requête, l'indexation porte sur un texte court et doit être exhaustive.

L'hypothèse d'une symétrie, si elle est, au travers de ces deux exemples, quelque peu prise en défaut, n'est pas, on le voit, réellement mise en question, malgré les impasses où elle mène invariablement.

Les impasses apparaissent clairement dès lors qu'il devient nécessaire de distinguer les objets, ne serait-ce que pour les mesurer, les évaluer[1], notamment parce que l'on ne sait pas exactement qui, de l'indexeur ou de l'utilisateur d'un système d'information[2], est censé dégager le « contenu », les informations d'un document.

En effet, dans l'hypothèse de la symétrie de l'indexation, la construction de l'information n'est le fait de personne : elle est le fruit d'une rencontre, presque fortuite, entre deux ensembles de mots élaborés pourtant par des acteurs différents, animés d'intentions différentes. Malgré ces différences, l'indistinction entre documents et requêtes peut continuer à se penser, comme le note Danièle Dubois, grâce à « la notion d'information [qui] a ainsi la vertu d'unifier la sémantique des divers objets tant matériels que matérialisés ou mentaux[3] ».

Quant à la rencontre et à la confrontation de ces différentes « informations » qui se donnent sous la même forme, quand elles ne sont pas purement et simplement ignorées, elles sont décrites comme relevant d'une « stratégie » qui ne concerne plus vraiment l'indexation : « L'essentiel dans un système documentaire est que la représentation des documents d'une part et la question d'autre part se rejoignent. Quant à définir à quel endroit se situe le point de jonction, cela est du ressort de la stratégie du système documentaire.[4] »

En évacuant la question du « lieu de rencontre » entre documents et requêtes, l'hypothèse de la symétrie apparaît comme ce qui permet de ne pas poser la question de la construction de l'information[5].

L'hypothèse de la symétrie, à l'œuvre dans les approches classiques de l'indexation, permet donc d'évacuer, sans justification véritable, les questions de la place et du rôle des différents acteurs – indexeurs et utilisateurs – de l'indexation ; autrement dit, dans le cadre d'une approche classique, ces questions ne relèvent pas, semble-t-il, d'une étude de l'indexation.

Si l'on s'en tient au discours classique, il paraît difficile de conduire une étude de l'indexation en tant qu'objet autonome, hors du cadre de la recherche

---

[1] Voir, sur ce point, Sparck-Jones (ed.) 1981, p. 214 : « It is difficult to valid access indexing correctness without retrieval. Retrieval however is not a real solution : why should a particular set of queries be used to test indexing ? No indexers or judges can foretell all future uses of documents ».

[2] Dans la suite de ce document, on emploiera uniquement le terme « utilisateur » pour faire référence à l'utilisateur d'un système d'information (automatisé ou pas).

[3] Dubois 1995, p. 89.

[4] Lallich-Boidin 1986, p. 162.

[5] Dubois 1995, p. 89 : « La recherche documentaire et plus généralement les sciences de l'information reposent sur une théorie de l'information qui, à la différence d'une théorie sémantique, ne pose ni la question de la production du sens dans un texte, par un auteur ou un utilisateur formulant sa requête, ni celle, symétrique, de l'interprétation ».

documentaire. On aurait néanmoins tort d'entériner une telle impossibilité sans préalablement examiner sur quoi elle repose.

L'analyse des présupposés du discours classique se révèle à ce titre éclairante. Elle permet de formuler les remarques suivantes :

- *l'indexation est définie sous l'angle d'une finalité spécifique – la recherche documentaire – donnée comme extérieure à l'indexation. L'approche est en cela essentiellement « instrumentale » et tend à introduire une circularité d'analyse ;*

- *suffit-il de définir autrement la recherche documentaire pour dégager l'indexation d'une telle circularité ? La tentative serait, semble-t-il, insuffisante : l'approche classique adopte implicitement un modèle de communication (le modèle mécaniste) qui contraint la recherche documentaire à n'être qu'une procédure d'appariement ;*

- *faut-il alors considérer que l'indexation et la recherche fonctionnent comme les deux faces d'un même processus qui, lui, pourrait constituer un objet de recherche ? Force est de constater que cet objet semble bien impalpable, mais surtout qu'il conduit à entériner l'hypothèse que la fabrication de l'information ne concerne pas en propre le traitement documentaire : voilà une conclusion qui remettrait en cause le projet même des sciences de l'information et de la communication.*

Parallèlement, l'examen des présupposés de l'approche classique de l'indexation permet de pointer les difficultés que rencontre une telle approche pour décrire précisément ce par quoi elle définit l'indexation : sa finalité. Le jeu de renvoi entre indexation et recherche, l'hypothèse opacifiante du « service à rendre », l'absence de prise en compte du rôle des différents acteurs, aboutissent chacun à repousser la question de la formulation précise et explicite de la finalité de l'indexation.

En quoi une approche non « instrumentale » de l'indexation peut-elle permettre de mieux l'appréhender ?

I.3 - Pour une approche non « instrumentale » de l'indexation

La difficulté de constituer l'indexation comme objet d'étude autonome apparaît principalement dans le cadre de définitions qui lient intimement indexation et recherche documentaires.

Or on peut montrer que ces définitions véhiculent implicitement des hypothèses relevant de modèles jugés désormais partiels ou insuffisants. Ces modèles présentent en outre le fâcheux inconvénient de ne pas offrir de réponses claires aux questions suivantes : quelle est la finalité de l'indexation ? Quelle est la « stratégie » à mettre en œuvre pour qu'un utilisateur ait accès à des documents ?

Il apparaît donc que l'on doit se dégager de ces définitions « instrumentales » ; mais apparaissent alors de nouvelles difficultés :

- d'une part, un paradoxe : peut-on étudier l'indexation en tant que pratique professionnelle finalisée tout en ne considérant pas cette finalité comme constitutive de la pratique ?

- d'autre part, une question : quelle est la finalité de l'indexation, si elle ne peut s'exprimer ni en termes de recherche documentaire ni en termes de « service » ?

Si l'on peut montrer que la finalité de l'indexation n'est pas à ce point ni spécifique ni « extérieure » à son mode de fonctionnement, alors une analyse de l'indexation comme objet d'étude autonome est possible.

Notre recherche conduit précisément à argumenter ces deux points : on pourra d'une part montrer que l'indexation peut se comprendre dans le champ plus vaste de la « diffusion des connaissances », où elle croise ce faisant d'autres pratiques professionnelles, comme celle de la vulgarisation scientifique[1]. Nous proposerons un rapprochement de ces deux pratiques du point de vue de leur finalité. D'autre part, la finalité de l'indexation se laisse déduire de l'usage de la langue qu'elle met en œuvre : sur ce point, l'indexation rejoint la problématique classique de la mise en relation entre les « mots » et les « choses », question à laquelle l'indexation propose une réponse « professionnelle », une réponse qui met en œuvre des moyens spécifiques à son champ d'exercice.

On fait donc l'hypothèse que l'on peut constituer l'indexation comme un objet d'étude autonome tout en rendant compte de la façon dont elle réalise son objectif : la finalité de l'indexation est alors inscrite dans le mode de fonctionnement même de l'indexation.

Cependant, à ce stade de la recherche, la possibilité d'une étude autonome de l'indexation ne peut être que présupposée. L'adoption de ce présupposé n'est pas sans conséquence.

Uniquement possible dans le cadre d'une approche non « instrumentale », la constitution de l'indexation comme objet d'étude autonome amène à modifier certaines des caractéristiques classiquement définitoires de l'indexation. Ces modifications portent essentiellement sur trois aspects :

- les objets de l'indexation : hors du cadre de l'hypothèse de la symétrie, l'indexation ne peut porter que sur les documents ; il ne sera pas question, dans cette recherche, de l'indexation des requêtes ;
- le processus de l'indexation : hors du cadre de l'hypothèse de service, l'indexation ne peut être comprise comme un simple transfert d'information ; elle sera plutôt appréhendée comme une opération, qui, à ce titre, produit et construit ses propres éléments (les documents et l'information, notamment) ;
- les outils de l'indexation : hors du modèle réductionniste de l'appariement, la problématique de l'indexation peut être dégagée de celle des langages documentaires ; on peut envisager d'autres moyens linguistiques par lesquels pourrait se réaliser l'indexation.

L'adoption d'une approche non « instrumentale » de l'indexation amène donc à évacuer, du champ de la recherche, l'indexation des requêtes, à mettre à distance le

---

[1] Voir, sur ce point, Jacobi [1987] qui propose de considérer les moyens de diffusion de la connaissance sous la forme d'un continuum : « La vulgarisation scientifique *s'inscrit dans un continuum de la diffusion de la science*, elle en est une des modalités. [...] La diffusion large, auprès d'un public indifférencié, par le moyen de rhétorique particulière n'est probablement qu'un cliché dénué de consistance. Dans les faits, c'est *une large panoplie de pratiques de socio-diffusion de la science* qu'il faudrait évoquer. » Jacobi 1987, respectivement p. 8 et p. 163. C'est nous qui soulignons.

rôle des langages documentaires et à considérer au sens fort la notion de « processus » en indexation.

Le premier paragraphe de ce chapitre a permis de mettre en valeur le caractère problématique du statut de l'indexation comme objet d'étude autonome. Notre recherche s'appuie sur le présupposé qu'un tel statut peut néanmoins être établi.

II - Définir la méthode d'analyse : approches théoriques de l'indexation

Compte tenu de l'absence de spécificité accordée à la seule indexation dans les discours classiques, on ne s'étonnera pas d'y trouver peu d'approches théoriques spécifiques. N'est-on pas alors amené à interroger la notion même d'approche théorique de l'indexation ? Peut-on distinguer, dans ce cadre, « théorie » et « fondement théorique » (II.1) ?

Dès lors que l'indexation ne semble pouvoir constituer l'objet d'une théorie mais qu'elle trouve ses fondements dans plusieurs théories, quelle approche théorique peut-on légitimement privilégier ? En quoi une approche linguistique des fondements théoriques de l'indexation peut-elle constituer un point de vue pertinent (II.2) ?

Enfin, si l'indexation peut se concevoir comme un objet d'étude autonome, peut-elle constituer, en soi, un objet d'étude pour la linguistique ? Quelle méthode d'analyse permet d'étudier l'indexation sous l'angle d'une problématique linguistique (II.3) ?

II.1 - Théories ou fondements théoriques de l'indexation ?

La possibilité de constituer une théorie de l'indexation est loin d'être partagée par tous les chercheurs du domaine. L'examen des productions scientifiques dessine principalement deux tendances :

- pour les uns, une théorie de l'indexation n'est purement et simplement pas envisageable[1] ;
- pour les autres, une théorie de l'indexation est envisageable, mais elle ne peut, en l'état actuel des connaissances, être élaborée[2].

On ne peut, à l'instar de Dachelet, que constater la grande absence des travaux théoriques sur l'indexation, absence qui contraste avec l'intense productivité des études concernant la recherche d'information : « Le travail qui s'accomplit en ce domaine [celui de la représentation du contenu des documents] est un travail de

[1] C'est, par exemple, la position de Varet 1995, p. 630 : « L'indexation est une question complexe parce qu'elle ne comporte pas de solution universelle. [...] Il n'existe pas même de théorie générale permettant d'en maîtriser intellectuellement le schéma ».

[2] Voir, par exemple, la position de Hutchins : « As our knowledge about the process of document representation remains inadequate, methods from AI and cognitive science are indeed called for : the process of document representation becomes an engineering problem ». Hutchins 1977 cité *in* Endres-Niggemeyer 1989, p. 231.

fond dont les résultats progressent lentement. Notre second thème : la recherche d'information, semble, par contraste, être l'objet d'une activité beaucoup plus intense et quasiment monopoliser les efforts de la recherche.[1] »

Comment interpréter cette faiblesse de la recherche ?

En examinant les arguments avancés par ceux qui dénient la possibilité d'une théorisation de l'indexation, on remarque que le principal argument mis en avant concerne l'utilité ou encore la pertinence d'une telle recherche[2], alors que, par contraste, les travaux sur la recherche documentaire semblent, eux, plus évidemment opératoires[3].

Il est clair, et on ne saurait s'en étonner, que ces arguments et ce choix d'axes de recherche s'inspirent directement et exclusivement d'une approche purement « instrumentale » de l'indexation, propre au modèle mécaniste de l'information dégagé précédemment.

Mais ce modèle, que l'on peut qualifier de « vicié » en ce qu'il est repose sur l'hypothèse opportuniste du « service », conduit en fait à une contradiction :

- d'un côté, l'indexation n'y est vue que comme une technique, et à ce titre, elle n'intéresse les chercheurs que par les résultats qu'elle fournit ; cependant, en tant que technique, l'indexation devrait pouvoir faire l'objet d'une théorie. Or c'est une théorie de la recherche documentaire qui est établie ;
- d'un autre côté, l'indexation, jugée trop complexe, est implicitement assimilée à une pratique et se pose à ce titre comme non théorisable.

Pour rendre plus claire la confusion de statut (pratique/technique) qui, à nos yeux, caractérise l'indexation dans les approches qui la tiennent pour non théorisable, il importe de préciser ce que l'on peut entendre par « technique » et par « pratique[4] ».

À la suite de Corinne Buisson[5], on reprend la distinction qu'Habermas[6] établit entre une technique et une pratique :

- une technique se comprend comme la version appliquée d'une théorie, elle est plus précisément « une application linéaire et univoque[7] » des lois ou des procédés que dégage une théorie ; c'est pourquoi ses résultats peuvent être tenus pour prédictibles et reproductibles à l'identique ;
- une pratique se réfère, elle, non pas à une théorie mais à un usage, à une norme ; elle tire son domaine de validité d'un système de valeurs, d'une

---

[1] Dachelet 1990, p. 1.
[2] Voir, par exemple, la position de Lancaster 1991, p. 28 : « A number of " theories " of indexing have been put forward [...] but these tend not to be true theories and they offer little practical help for indexer ».
[3] Voir, par exemple, cette remarque de Endres-Niggemeyer [1989, p. 230] : « This is indeed a very common attitude among information scientists that we do not need to know how indexers arrive at a particular description of the contents of a document ; all that matter is whether it enables users to find the document when required ».
[4] Cette distinction nous est apparue essentielle à la lecture de Buisson 1995, notamment la problématique qu'elle pose p. 122 : « Doit-on parler d'une pratique de l'indexation ou d'une technique ? L'emploi de l'un ou l'autre terme n'est pas sans incidence sur la signification que l'on attribue à l'opération et sur le rôle que l'on assigne au documentaliste qui indexe ».
[5] Buisson 1995, p. 121-128.
[6] Habermas 1973 [1968].
[7] *Ibid.*, p. XXXV.

« idéologie » entendue au sens de Boudon[1] : ses productions sont évaluées à cette aune.

Ces deux notions succinctement posées, on peut reformuler l'impasse et la contradiction auxquelles mène le modèle mécaniste de l'information : il incite à concevoir l'indexation comme une technique mais ne donne pas les moyens de restituer la théorie dont elle est censée être l'application ; comme la théorisation de l'indexation s'avère impossible, on pose alors implicitement que l'indexation est une pratique, et donc qu'elle ne peut faire l'objet d'aucune théorie.

La mise en valeur de cette confusion de statut peut suggérer une réponse à notre interrogation sur la rareté, voire l'inexistence, de théories de l'indexation, alors que pourtant certains la considèrent possible : l'indexation relèverait moins d'une technique, supposant une théorie, que d'une pratique, reposant sur un système de valeurs, une idéologie. Dès lors, il faut renoncer à toute tentative de théorisation et considérer pleinement l'indexation sous l'angle d'une pratique[2]. Là encore, seules des approches qui s'écartent du modèle mécaniste peuvent tenir l'hypothèse de l'indexation comme pratique ; notre adoption d'une approche non « instrumentale » nous permet de la penser comme telle.

Notre recherche posera donc comme hypothèse que l'indexation constitue une pratique.

L'indexation considérée comme une pratique impose un type de relation particulier aux champs théoriques. Cette relation se joue sur le mode de la *représentation** : une pratique professionnelle comme celle de l'indexation repose, de façon implicite, sur une certaine représentation d'une ou plutôt de plusieurs théories (une théorie du langage, une théorie de la cognition, une théorie de la communication, etc.). C'est en ce sens que, selon nous, une étude scientifique de l'indexation ne peut être qu'une étude de ses fondements théoriques, c'est-à-dire qu'une étude de la représentation des théories sur lesquelles implicitement elle se fonde. Mais une étude nécessairement critique, dans la mesure où toutes les représentations théoriques sous-jacentes à l'indexation ne peuvent être comprises comme des fondements théoriques, c'est-à-dire comme des fondements valides du point de vue d'une théorie constituée.

L'étude des fondements théoriques de l'indexation se dédouble donc nécessairement en deux volets :

- une analyse descriptive des représentations théoriques sous-jacentes à la pratique d'indexation ;
- une analyse critique évaluant les distorsions (éventuelles) qui s'observent entre théorie scientifique et théorie représentée en indexation.

Pour mener cette étude des fondements théoriques de l'indexation ainsi entendue, devra être élaborée une méthodologie d'étude qui permette de confronter les deux niveaux en présence : le niveau relatif à la théorie sollicitée en indexation et le

---

[1] Boudon 1986, p. 82 : l'idéologie peut être vue comme une « interprétation significative du monde échappant au critère de la vérité et de l'erreur et pouvant être expliquée à partir de l'environnement social de l'acteur ».

[2] Comme le note Corinne Buisson [1995, p. 121-123], il est rare que l'indexation soit posée, jusque dans ses ultimes implications, comme une pratique : la notion de « techniques documentaires » reste dominante.

niveau relatif à la représentation qui est faite de cette théorie dans le cadre et pour les besoins d'une pratique.

Avant d'aborder la question d'une telle méthodologie, nous devons déterminer le type de théorie à partir de laquelle la pratique d'indexation construit ses représentations. À la suite d'autres auteurs[1], nous faisons l'hypothèse que les fondements théoriques de l'indexation, comme l'ensemble des objets d'étude des sciences de l'information et de la communication, relèvent de plusieurs champs disciplinaires[2]. L'un de ces champs est celui des sciences du langage* : est-ce le plus essentiel, le plus déterminant ? Nous l'ignorons[3] ; mais il importe de noter que le champ de la linguistique est celui qui apparaît le plus visiblement sollicité dans la pratique de l'indexation (voir ci-après, § II.2).

Qu'est-ce que les sciences du langage nous permettent de voir de l'indexation et qu'est-ce qu'elles ne nous permettent pas d'atteindre ? Si seule une approche interdisciplinaire pourrait permettre de le dire précisément, nous essaierons néanmoins de circonscrire les objets de l'indexation dont une approche linguistique peut rendre compte (§ II.3).

II.2 - Pour une approche linguistique des fondements théoriques de l'indexation

Si elle peut paraître relativement « naturelle[4] », une approche linguistique de l'indexation ne présente cependant aucun caractère d'évidence. En effet, ce n'est pas parce que l'indexation manipule, entre autres, des objets de nature linguistique (textes des documents, mots des langages documentaires par exemple) qu'elle manipule des objets de la linguistique. En effet, sans faire une épistémologie des sciences du langage, on doit relever que la linguistique, en tant que discipline scientifique, d'une part n'est pas une (il y a plusieurs types de linguistique[5]) et, d'autre part, qu'elle n'a pas pour ambition de rendre compte de tout le langage ni de toutes ses réalisations[6]. De façon générale, on peut dire que l'objet de la linguistique[7] est la description de ce qui constitue une langue, c'est-à-dire, entre autres, la mise au jour de propriétés qui permettent de distinguer une langue d'une non-langue. Dans ce cadre, les objets que manipulent généralement les linguistiques sont des objets qui ne peuvent, à strictement parler, être traités que par elles et qui n'intéressent qu'elles[8].

Cependant, du point de vue qui est le nôtre (celui de l'analyse des représentations de la langue dans la pratique d'indexation), la linguistique, entendue comme « un

---

[1] Miège par exemple 1993-1994.
[2] Voir, sur ce point, Le Coadic 1994.
[3] Blair [1990, p. 122] fait l'hypothèse que toute approche de l'indexation repose sur un modèle du langage : « Any theory of indexing or document representation presupposes a theory of language and meaning ».
[4] Par exemple Menon 1988, p. 165 : « Les textes, en tant qu'objets langagiers, posent avant tout – et qui s'en étonnera – des problèmes linguistiques ».
[5] Milner 1978, p. 11.
[6] Milner 1989, p. 38-50.
[7] Voir Milner *Ibid.*, p. 43 et suiv.
[8] Milner fait observer (*ibid.*, p. 34) que « dès qu'elle dépasse la banalité, une proposition linguistique concerne peu de données à la fois et elle fait apparaître généralement ce que l'opinion courante tiendrait pour des détails ».

ensemble de postulats sur la langue (ou le langage) et de méthodologies de description[1] », nous paraît pertinente, et ce à deux niveaux :

- au niveau des descriptions des faits de langue : les propriétés que la linguistique dégage à partir de l'étude de ses propres objets peuvent, moyennant une décontextualisation et une généralisation, expliquer le fonctionnement d'un certain nombre de mécanismes à l'œuvre dans l'indexation, mécanismes peu visibles hors du prisme de l'analyse linguistique ;

- au niveau des postulats sur la langue* et le langage* : la théorie linguistique, si elle n'est pas la seule à tenir des propositions générales sur la langue et le langage, fournit l'un des cadres possibles pour discuter la conception du langage sous-tendue par la pratique d'indexation ; à ce titre, on pourra confronter la conception du langage telle que la véhicule l'indexation à la conception du langage postulée par les modèles linguistiques.

On peut, à titre d'exemple, citer quelques caractéristiques qui, issues de la pratique d'indexation, peuvent être appréhendées dans le cadre d'une approche linguistique, à l'un ou à l'autre des deux niveaux dégagés.

(i) Au niveau des descriptions des faits de langue

On liste ci-dessous, sans souci d'exhaustivité ni de détails, quelques particularités de l'indexation ou des langages documentaires qui recoupent des descriptions de faits de langue menées en linguistique.

Convention d'écriture ou propriété des unités lexicales ?

Les langages documentaires utilisés en indexation sont constitués exclusivement de noms[2]. Dans les termes de l'approche classique, il s'agit simplement de se donner là une convention d'écriture[3]. Or, sur un plan linguistique, les unités lexicales de catégorie nominale sont des unités linguistiques qui présentent la propriété de pouvoir, en discours, référer à un objet. Nous reviendrons largement sur cette propriété linguistique qui fait apparaître la notion cruciale en indexation de référence*, cette notion ne pouvant apparaître en tant que telle si l'on se place du seul point de vue de la pratique.

Synonymie linguistique ou synonymie documentaire ?

La notion de synonymie qui relève du champ de la linguistique est utilisée en indexation dans une acception qui se veut particulière : on parle en effet de synonymie documentaire, que l'on oppose à synonymie linguistique, lorsque l'on veut décrire la possibilité de rapprocher des termes alors « que leur signification

---

[1] Marandin 1979, p. 18.

[2] Au sens de l'anglais *noun*, catégorie grammaticale, qui s'oppose à *name*, le « nom » du référent : le terme français ne rend pas cette distinction.

[3] Voir, par exemple, la norme AFNOR Z 47-100 (1981), p. 187 : « La forme grammaticale du descripteur doit se conformer aux *règles suivantes : forme substantive* [...] », ou encore Chaumier 1978, p. 31 : « Les descripteurs sont des termes normalisés, les *règles d'écriture* d'un descripteur étant les suivantes : *forme substantive* ». C'est nous qui soulignons.

sémantique [est] différente[1] ». Une description linguistique permet justement de montrer que la synonymie ne joue pas sur le sens des unités mais sur leur référent, ouvrant ici une nouvelle problématique pour l'étude de la référence en indexation : celle du rapport entre le sens et la référence, nous y reviendrons. Là encore cet enjeu ne peut apparaître clairement si l'on reste au niveau du seul discours classique.

Les « formes nouvelles de composition nominale[2] »

Dans son article de 1966, Benveniste relevait l'« extension considérable » d'un type de termes « construit[s] sur un modèle qui n'est plus celui de la composition classique » : il s'agit de ce qu'il nomme les synapsies, qui sont du type « modulation de fréquence », « avion à réaction[3] ». Benveniste prédit que ce type de termes « sera la formation de base dans les nomenclatures techniques[4] ». Depuis, des études ont été menées en linguistique qui portent sur la description de telles unités : nous évoquerons l'une d'entre elles. On peut relever qu'à la même époque (dans les années 1970) apparaissent, dans les thésaurus, de nouveaux types de descripteur ayant précisément pour particularité d'être constitués de plusieurs « mots » (de type « contrôleur de gestion », « énergie géothermique[5] »). Dans le discours classique, l'introduction de telles expressions ne se justifie que par rapport aux difficultés que rencontre une interrogation par unitermes[6]. Par le biais de la description linguistique de la synapsie, on pourra montrer que cette apparition d'une nouvelle forme de descripteur peut relever d'une tout autre raison que celle de la seule performance en matière de recherche documentaire : une unité telle que la synapsie révèle en effet des propriétés qui intéressent au premier chef l'indexation elle-même et non simplement la recherche documentaire.

(ii) Au niveau des postulats sur la langue et le langage

Si l'on ne prétend pas que seules les théories linguistiques sont à même de discuter les postulats que l'on peut tenir sur la langue et sur le langage, on tient que la conception du langage qui les constitue comme science peut fournir un référentiel pertinent pour penser les représentations du langage qui sous-tendent les pratiques d'indexation, et notamment :

La représentation de la langue comme une nomenclature

C'est l'adoption de cette représentation particulière du langage qui, semble-t-il, autorise la pratique documentaire à créer un langage artificiel à partir de la langue « naturelle ». Sur la base d'une critique de la vision de la langue comme nomenclature, la linguistique invite à repenser cette question et à substituer à la

---

[1] Chaumier 1978, p. 33.
[2] Titre d'un article de Benveniste paru en 1966 et repris dans le chapitre XII de Benveniste 1974, p. 163-176.
[3] Exemples repris de Benveniste. *Ibid.*, p. 172.
[4] *Id.*
[5] Exemples repris de Chaumier 1978, p. 23.
[6] Chaumier, *id.* : « Devant les problèmes posés par l'utilisation de la post-coordination totale sur les unitermes au moment de l'indexation, un certain degré de précoordination se fit jour dans les thésaurus avec l'emploi des descripteurs. Il s'agit ici de précoordination au niveau du concept. C'est ainsi que les thésaurus admirent des expressions précoordonnées telles que "contrôleur de gestion", "énergie géothermique" composées chacune de deux unitermes ».

notion de création de langages documentaires celle de création d'une utilisation documentaire de la langue « naturelle[1] ».

La représentation de la relation entre les mots et les choses

Cette relation entre les mots et les choses, dite aussi relation de référence, est au cœur des problématiques de la linguistique. L'une des façons de la problématiser peut être : « comment le langage parvient-il à parler du réel ?[2] ». Par contraste, la relation référentielle n'est qu'implicitement présente dans le discours classique : cependant, nous pourrons observer que la problématique de la référence, une fois posée, apparaît en fait dans l'indexation sous des formes variées et qu'ainsi mise en valeur, elle permet de clarifier une partie du fonctionnement et de la finalité de l'indexation.

La question du sens

Là encore, et il n'est pas superflu de le noter, la question du sens n'est pas explicitement posée en indexation[3]. En revanche, le discours classique manipule largement la notion de « contenu » (d'un document, d'une requête). Cette terminologie n'est pas neutre, comme nous tenterons de le montrer : elle suppose une certaine conception du sens*, que les approches linguistiques, supposant souvent en des termes différents qu'il y a de la signification* dans et par la langue, nous permettront de mettre au jour.

Si ces quelques exemples peuvent montrer que la référence à la linguistique en tant que science est bien à l'œuvre dans la pratique de l'indexation, du moins au niveau implicite des représentations, par quelle méthode mener une étude qui permette à la fois de :

- *montrer que ces représentations linguistiques peuvent être, pour certaines d'entre elles, considérées comme des fondements théoriques de l'indexation ;*
- *maintenir distincts les deux types d'objet que nous aurons à manipuler – les objets de la linguistique et les objets de la pratique documentaire ?*

La principale difficulté tient à ce que, la linguistique n'ayant rien de particulier à dire sur l'indexation, nous serons nécessairement amenée à procéder à des déplacements (sous forme de décontextualisations et/ou des généralisations) qui introduiront nécessairement des décalages de niveaux : le niveau de l'analyse linguistique n'est pas celui qui est directement pertinent pour l'indexation. Ainsi, traiterons-nous, par exemple, de la référence à partir d'études linguistiques ayant pour objet l'anaphore ; de même aborderons-nous la question de la signification lexicale par le biais de l'étude morphologique d'un type de mots particuliers, les mots construits*, etc.*

---

[1] Milner 1989, p. 35 : « En fait, le problème doit être posé autrement : en tant que la science les saisit, les langues et le langage ne sont pas des matières réalisées ; ce sont plutôt les *lois* qui régissent ces "matières". Inversement, les techniques de la langue n'ont pas pour fin de produire de nouvelles entités de langue (de nouveaux mots, de nouvelles structures, de nouvelles langues, etc.), mais de nouveaux objets où les langues *telles qu'elles sont* interviennent : ce ne sont donc pas à proprement parler les langues qui sont visées, mais les réalisations de langue - textes, messages, slogans, discours, etc. ».

[2] On reprend ici une formulation proposée dans Kleiber 1981, p. 11.

[3] Dubois 1995, *supra* (note 26).

En outre, et ce second point constitue une autre limite importante de cette recherche, nous ne pourrons nous permettre de discuter la pertinence des descriptions linguistiques sollicitées, pouvant ainsi donner l'impression de ne retenir des linguistiques que la description de faits qui concerne de façon opportuniste notre objet.

On peut essayer de réduire l'incidence de ces deux limites en disposant d'une méthode d'analyse qui permette de donner un statut à ce type d'analyse du décalage entre modèle de langue et modèle d'usage de la langue. On attend de la méthode décrite ci-dessous qu'elle constitue un garde-fou évitant de céder à la tentation de plaquer une description linguistique sur des objets documentaires. On attend également de cette méthode qu'elle puisse accueillir d'autres types d'approches linguistiques que celles adoptées dans cette recherche : on espère disposer là d'un cadre qui permette aux modèles linguistiques que nous retiendrons de pouvoir être discutés dans le cadre particulier des problématiques de l'indexation.

II.3 - Une méthode d'analyse du décalage

Nous présentons ci-après, en nous appuyant sur Berrendonner et Reichler-Béguelin [1989], une méthode d'analyse valable pour tout type de démarche qui cherche à confronter les propositions d'un modèle théorique aux réalisations d'une pratique qui paraît s'inspirer de ses objets et/ou de ses présupposés (A).

Nous préciserons ensuite la forme que prend cette méthode appliquée à notre approche linguistique de l'indexation. Dans le cadre de cette méthode, nous identifierons les objectifs que notre recherche peut se donner et les objets sur lesquels elle peut porter (B).

Nous aborderons pour finir à la fois les contours et les limites que suppose notre recherche (C).

A - Présentation de la méthode d'analyse du décalage

La perspective dans laquelle Berrendonner et Reichler-Béguelin [1989] étudient les pratiques de segmentation d'un texte (en lettres, mots, phrases) s'inscrit dans un cadre méthodologique plus global qui distingue les « représentations formelles » d'une part et les « catégorisations pratiques » d'autre part.

Les représentations formelles sont celles que produit la science, les catégorisations pratiques sont celles que produisent les pratiques : d'un côté, des représentations établies selon des critères scientifiques ; de l'autre, des catégorisations régies par des contraintes utilitaires. Les deux types d'entité se distinguent sur plus d'un point : homogénéité *versus* hétérogénéité des propriétés, objet à un *versus* plusieurs critère(s), indéformabilité *versus* adaptabilité des classes d'objets. C'est pourquoi il est courant de noter que « toutes les catégorisations pratiques diffèrent des représentations produites par la science : almanach vs météorologie, classification sociale des aliments vs diététique[1] ».

Peut-on aller plus loin que la seule notification de cette différence ? Et, si oui, comment procéder ? Quelle méthode élaborer pour étudier les décalages entre

[1] Berrendonner et Reichler-Béguelin 1989, p. 106.

représentations formelles et catégorisations pratiques, sans les exclure ni les dissoudre les unes aux autres ? Berrendonner et Reichler-Béguelin adoptent une méthode qui repose sur la confrontation non des entités elles-mêmes mais des *modèles* dans lesquels elles s'intègrent ; ces modèles ne sont pas tout à fait de même nature.

En effet, si la constitution d'un modèle répond à une exigence scientifique dans le cas des représentations formelles, elle répond à une nécessité d'usage dans le cas des catégorisations pratiques. La nature de chacun des deux modèles est donc foncièrement différente : « En matière d'interactions sociales, il importe que les divers usagers d'un même système disposent d'emblée d'un minimum de *représentations communes*, faute de quoi le temps d'agir se perd à négocier un accord des partenaires sur la « façon de voir les choses ». Ce besoin de fonder la coopération sur une base de représentations communes explique que les catégorisations pratiques empruntent volontiers à une *doxa*, à une tradition collective, ou à un fonds idéologique majoritaire des schèmes cognitifs stéréotypés, qui bénéficient de l'immédiateté de l'évidence. Ceux-ci sont alors ressentis, et généralement dépeints, comme un corps de normes sociales imposées aux individus.[1] »

On appellera la *doxa* d'une pratique « modèle d'utilisation », que l'on opposera au modèle scientifique, nommé alors « modèle de fonctionnement[2] ».

La méthode permet ainsi d'étudier des décalages entre modèles, c'est-à-dire des décalages entre modes de saisie des objets. C'est pourquoi ce type d'étude repose essentiellement sur l'analyse des zones de « tension » : « Lorsqu'un modèle formel et une catégorisation pratique se disputent le même objet, il s'établit entre eux une tension dialectique : chacun des deux tend à réduire l'autre par absorption de certains de ses schémas.[3] »

B - Méthode d'analyse du décalage linguistique en indexation

Berrendonner et Reichler-Béguelin adoptent une stratégie d'analyse en termes de décalage pour atteindre ce que les sujets parlants considèrent comme des mots, des phrases, etc., et non pour vérifier la compatibilité de leur vision de linguistes avec celle des alphabètes lambda[4]. Leur approche repose sur l'hypothèse que « les unités non formelles, créées, récupérées ou bricolées à des fins techniques, portent la trace d'un mode de structuration pragmatique du langage obéissant à des règles propres, en vertu de finalités spécifiques[5] ». De la même façon, nous pensons que les objets spécifiques à l'indexation (descripteur, langage documentaire par exemple) portent les traces d'une certaine vision de la langue en même temps que celles d'un certain usage de la langue.

C'est pourquoi la méthode d'analyse du décalage nous paraît pouvoir être utilisée : d'une part, pour dégager les représentations linguistiques sous-jacentes à la pratique d'indexation ; d'autre part, pour montrer comment ces représentations peuvent, pour certaines d'entre elles, se constituer en fondements théoriques. Le passage du

---

[1] Berrendonner et Reichler-Béguelin 1989, p. 109. C'est nous qui soulignons.
[2] Berrendonner et Reichler-Béguelin reprennent ici une distinction proposée par Amalberti [1987].
[3] Berrendonner et Reichler-Béguelin 1989, p. 111.
[4] *Ibid.*, p. 103.
[5] *Ibid.*, p. 100.

niveau des représentations à celui des fondements n'est en effet pas automatique : seules quelques représentations linguistiques à l'œuvre en indexation relèvent d'un modèle de fonctionnement formel de la langue et peuvent à ce titre prétendre au statut de fondements théoriques (*i.e.* fondements du point de vue de la théorie linguistique). Dans d'autres cas, les représentations linguistiques sous-jacentes en indexation se révèlent être des représentations *ad hoc* établies *a posteriori* par la pratique, qui n'ont plus alors qu'un lointain rapport avec un quelconque modèle de la langue.

En effet, si les représentations non formelles de la langue en indexation se trouvent régulièrement en conflit avec les représentations formelles des linguistes, il arrive aussi qu'elles les « absorbent » complètement. Dans les termes de Berrendonner et Reichler-Béguelin, il est en effet inévitable que, se disputant les « mêmes objets », les deux types de modèle cherchent à s'absorber et à se confondre ; un modèle d'utilisation de la langue finit par se donner pour un modèle de fonctionnement, alors que le modèle d'utilisation devrait uniquement « utiliser », c'est-à-dire, étendre, déformer, etc., les éléments d'un modèle formel : « La récurrence obstinée de ces polyvalences où sont mis en jeu deux ou plusieurs niveaux d'articulation du langage théoriquement distincts, est à mettre au compte d'une logique utilitaire construisant et adaptant ses unités en fonction de pertinences opératoires multiples. Ces chevauchements sont intrinsèquement liés à l'élaboration spontanée des outillages graphiques et *par contrecoup, ils conditionnent forcément la perception non formelle que les sujets ont des unités de leur langue.*[1] »

Autrement dit, le modèle d'utilisation de la langue outre qu'il « pervertit[2] » régulièrement, pour ses besoins propres, les représentations formelles, se place en constante concurrence avec le modèle de fonctionnement de la langue. C'est par cette double tension que nous pourrons montrer comment se constituent les représentations linguistiques en indexation et comment seulement certaines d'entre elles peuvent fonctionner comme des fondements théoriques.

Par le biais de la méthode d'analyse du décalage, nous chercherons donc à :

(i) dégager, à partir de la *doxa*[3] linguistique en indexation, le modèle d'utilisation de la langue mis en œuvre, ainsi que le modèle de fonctionnement de la langue auquel il se réfère implicitement : à cette étape, on cherche à dégager les représentations de la langue qui sous-tendent la pratique de l'indexation ;

(ii) confronter le modèle de fonctionnement implicite en indexation au modèle de fonctionnement explicite en linguistique : à cette étape, on cherche à évaluer la validité des représentations de la langue véhiculées en indexation au regard des hypothèses linguistiques. Les représentations de la langue en indexation

---

[1] Berrendonner et Reichler-Béguelin 1989, p. 102 (c'est nous qui soulignons).
[2] *Ibid.*, p. 111 : « Les catégorisations pratiques incorporent chroniquement - en les pervertissant - des notions scientifiques vulgarisées ».
[3] On rappelle que la notion de « doxa » a été proposée par Aristote dans le cadre de ce qu'on a pu appeler une théorie des lieux du langage. La *doxa* constitue le second de ces lieux, c'est le lieu « commun » à tous les sujets parlants qui, à ce titre, peut être vu comme un « tissu de conjectures, d'usages habituels, de comportements les plus ordinaires, de discours vraisemblables », Cauquelin 1990, p. 66-67. Sa principale caractéristique est de constituer la « matière première de l'entente, voire de la concorde, parce que sans elle il n'y aurait rien à partager, rien non plus à préciser, à distribuer selon les lieux », *Ibid.*, p. 34. C'est pourquoi la *doxa* n'intéresse pas directement par la validité des propos qui s'y tiennent.

sont ou pas conformes aux représentations formelles des linguistes : si elles le sont, elles sont considérées comme constituant des fondements théoriques ; si elles ne le sont pas, on procède à un déplacement de modèle de fonctionnement ;

(iii) substituer, aux représentations linguistiques de l'indexation non valides sur le plan linguistique, des représentations linguistiques valides, qui constituent alors des fondements théoriques pour l'indexation ; dans le cadre de ces nouvelles représentations, on reformule le modèle d'utilisation de la langue en indexation. À cette étape, on propose une redéfinition de l'indexation (ou plutôt de certains aspects de l'indexation) où les références et les emprunts à la linguistique sont conformes à un modèle théorique de la langue.

On peut schématiser la méthode que nous adopterons pour constituer les fondements de l'indexation du point de vue de la théorie linguistique de la façon suivante[1] :

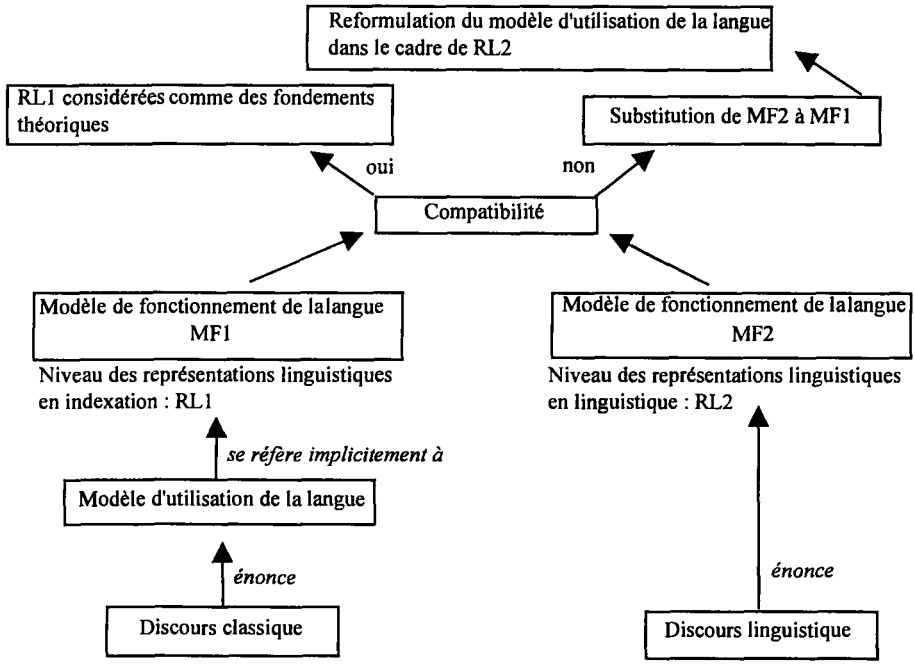


Figure 2 – La méthode d'analyse du décalage

Il importe de noter que l'adoption d'une telle méthode favorise, d'une certaine façon, le « détournement » des concepts scientifiques. Sur ce point, il est clair qu'une étude des fondements théoriques de l'indexation telle que nous la proposons (par opposition à une étude qui chercherait à constituer une théorie de l'indexation par exemple) ne permet pas de sortir à proprement parler du « bricolage théorique »

[1] Il nous semble que la méthode adoptée ici peut être aussi utilisée pour la constitution des fondements de l'indexation du point de vue d'autres théories (cognitive, sociologique, etc.).

qui caractérise la pratique d'indexation. L'objectif consiste simplement à se doter d'un référentiel (ici une théorie linguistique) qui, en étant étranger au domaine documentaire lui-même, permette de sortir du discours de la pratique sur elle-même, et, de ce fait, d'objectiver certains faits d'indexation. Le référentiel proprement documentaire, le discours classique, en ne s'inscrivant dans le cadre d'aucune théorie, ne peut permettre de dépasser la complexité d'une pratique et de « désintriquer » les différents phénomènes en jeu. Par le référentiel linguistique, on pourra faire apparaître une partie des mécanismes de l'indexation : ceux qui reposent sur les propriétés de la langue.

Pour rendre complète la présentation de la méthode d'analyse que nous suivrons dans cette recherche, nous devons déterminer les objets de l'indexation qui peuvent être étudiés dans le cadre proposé. Nous en avons dégagé cinq, susceptibles de nous permettre d'énoncer les fondements de l'indexation du point de vue d'une théorie linguistique. Les objets de l'indexation ne correspondant pas aux objets de la linguistique, nous présentons dans le tableau suivant, au regard des objets documentaires retenus pour étude, les objets formellement constitués par les linguistiques, qui nous permettront d'avancer dans notre recherche :

Objets de l'indexation	**Objets formels**
descripteur	- terme textuel - groupe nominal - synapsie
document	- référent - énoncé
information	- objet de discours - thème discursif
langage documentaire	- discours [documentaire]
collection documentaire	- formation discursive

Les objets de l'indexation du tableau ci-dessus ne seront étudiés ni successivement ni séparément, ceci pour deux raisons : d'une part, ils mettent souvent en jeu les mêmes problématiques linguistiques (la référence par exemple pour l'étude du descripteur et du document), d'autre part, ils sont fortement liés entre eux (document et collection documentaire par exemple). Ces différents objets seront plutôt étudiés selon le point de vue qui rend visibles les « lieux de tension » entre les deux modèles formel et non formel de la langue : le lexique, la référence, le discours. Seul le descripteur, parce qu'il permet de rendre compte globalement de l'indexation fera l'objet d'un chapitre particulier, dans lequel nous pourrons synthétiser l'ensemble de nos propositions.

C - Limites et contours de la recherche

Le cadre méthodologique que nous nous proposons de suivre ne permet évidemment pas de traiter tous les aspects de l'indexation : comme nous l'avons vu, on ne peut constituer un modèle d'utilisation qu'à partir d'une seule théorie à la fois. Or, nous avons épousé l'hypothèse que la pratique d'indexation, comme la plupart des pratiques, devait relever de plusieurs champs disciplinaires : à ce titre, elle met en jeu des représentations théoriques de nature différente. Par ailleurs, il n'est pas

exclu qu'une même représentation puisse relever de modèles formels différents (par exemple, la notion de « concept » en indexation doit sans doute reposer à la fois sur des représentations linguistiques et sur des représentations cognitives).

Il importe donc d'essayer de formuler les limites qui, à plusieurs niveaux, sont impliquées par le cadre méthodologique retenu :

- limite relative à la couverture de l'objet d'étude : dans l'approche linguistique de l'indexation que nous avons adoptée, nous ne pouvons étudier que des objets documentaires où la matière linguistique se trouve réalisée[1], c'est-à-dire essentiellement les descripteurs, les documents et les collections documentaires, les langages documentaires et l'information[2]. Il y a donc forcément de nombreux aspects de l'indexation dont cette recherche ne pourra rendre compte[3] ;

- limite relative à la couverture des objets étudiés : les objets empiriques de l'indexation que nous avons retenus pour notre étude mettent en œuvre, de façon évidente pour certains d'entre eux (comme la notion d'information), plusieurs dimensions et relèvent à ce titre de modèles formels différents. Il est, en l'état actuel, délicat de préciser la couverture descriptive qu'offre, pour un objet donné, une approche linguistique. En conclusion de cette recherche, nous essayerons d'indiquer, au moins pour les objets documentaires que nous aurons étudiés, les autres représentations théoriques qui nous semblent pouvoir être en jeu en indexation.

Les contours de cette recherche restent délicats à déterminer : on distingue mal ce que l'approche linguistique est à même de traiter et ce dont elle ne peut rendre compte. Nous pensons qu'une telle difficulté, liée en grande partie à notre approche de l'indexation, est par ailleurs amplifiée par la « nature » même du langage : s'il est clair que l'on peut, en plus d'un point d'une pratique, relever « des interventions en langue naturelle[4] », il n'est pas toujours évident de décider celles qui relèvent, « en droit », d'une approche linguistique et celles qui n'en relèvent pas. Sans doute aurons-nous laissé de côté, dans cette recherche, des aspects de l'indexation qu'une linguistique pourrait permettre de traiter, tandis que nous aurons parallèlement préjugé des fondements linguistiques de tels autres aspects de l'indexation. Là encore, il nous semble que la méthode d'analyse proposée permet d'introduire des modifications (par exemple dans les objets d'étude) sans que la démarche elle-même ne soit remise en cause.

Cependant – et ce sera là la dernière précision que nous apporterons à la méthode d'analyse adoptée –, ce n'est pas parce « l'intervention en langue naturelle », pour reprendre les termes de Milner, n'est pas immédiate que le modèle linguistique n'est plus approprié à la description des faits d'indexation. Sur ce point, on tiendra que

---

[1] Des « réalisations de langue » dans la terminologie de Milner 1989, p. 35.
[2] Pour ce dernier objet, plus que pour les autres, la réalisation de la matière linguistique n'est pas évidente et l'hypothèse de la représentation linguistique que nous ferons (sous la notion d'« objet de discours ») sera alors ici plus forte qu'ailleurs. Nous espérons pouvoir montrer que cette hypothèse n'est cependant pas dénuée d'intérêt.
[3] Par exemple, tous les aspects qui concernent la façon dont les indexeurs indexent ; sur ce point, on peut se reporter à l'analyse menée dans le cadre des sciences cognitives par Bertrand 1993.
[4] Milner 1989, p. 32.

notre approche linguistique permet de traiter l'indexation de tous les types de documents, que ceux-ci soient de nature textuelle ou non[1].

En effet, comme cette recherche se donne pour objectif de le montrer, la pratique de l'indexation met toujours en œuvre des discours, quel que soit le type d'objet manipulé (texte, image, son, etc.). Ces discours sont soit directement liés au tissu textuel du document à analyser : ils sont alors particulièrement peu visibles. Ces discours sont, dans le cas des objets non textuels, créés spécifiquement par et/ou pour l'indexation : ils offrent alors une meilleure visibilité. En effet, un objet documentaire à indexer est toujours accompagné d'un texte[2] : la présence d'un texte est une condition pour qu'un document constitue un objet documentaire[3]. Si ce texte n'existe pas, il faut le créer[4]. Nous aurons à expliquer cette présence obligatoire du texte dans l'objet à indexer.

Que les documents non textuels soient pris en compte dans notre approche linguistique de l'indexation ne signifie pas, pour autant, que nous ne tenons pas compte de la spécificité de ce type de documents ; tout au contraire. C'est parce que notre cadre d'analyse est un cadre linguistique que nous pouvons poser ces différences. En effet, dans le cadre d'une approche sémiologique[5] telle que Benveniste[6] a pu la définir, texte et image (par exemple) constituent deux modes sémiotiques différents, non réductibles l'un à l'autre[7]. Le texte n'est donc pas, avec l'image, en relation de commentaire ou d'interprétation.

Nous aurons donc à déterminer le type de relation que le texte établit avec l'image en indexation mais aussi l'utilité, pour l'indexation, de disposer d'un texte qui ne renvoie pas au « contenu » de l'image. Cet aspect nous intéresse au plus haut point puisqu'il met au jour un « paradoxe » : l'indexation porte, dans ce cas, non sur l'image elle-même mais sur le texte qui l'accompagne (sans l'interpréter), alors même que l'indexation est supposée réaliser une « analyse du contenu » du document (ici un document iconographique). On comprend donc que, dans le cadre

---

[1] L'ensemble des objets susceptibles d'être indexés est très divers et mal défini. On suppose que l'indexation est en principe exercée au moins sur l'ensemble des productions éditoriales qui font l'objet d'un dépôt légal. La loi n° 92-546 du 20 juin 1992 relative au dépôt légal français circonscrit le type de productions concernées. Sont ainsi candidats à l'indexation : les documents imprimés, graphiques, photographiques, sonores, audiovisuels, multimédias ; les progiciels, les bases de données, les systèmes experts et les autres produits de l'intelligence artificielle.

[2] Le cas exemplaire est celui du document iconographique.

[3] Voir la norme de catalogage des images fixes (Norme AFNOR Z 44-077, 1995) qui stipule que « l'image est toujours accompagnée d'un texte. [...] Le texte d'accompagnement est le document historiquement attesté par lequel l'image est identifiée ». Voir aussi Moles : « Une image sans légende de la part de son fournisseur ou producteur n'a pas - en principe et en général - droit d'accès à l'iconothèque », Moles cité *in* Le Guern Odile 1989, p. 427.

[4] Voir par exemple Escarpit 1991, p. 161 : « On notera que le documentaliste part toujours d'un texte. Si ce texte n'existe pas, il faut d'abord le constituer. Nous avons vu par exemple que l'analyse directe de l'image pose des problèmes insurmontables. Toute analyse documentaire d'une image doit donc commencer par la production d'un discours descriptif de l'image et la notation de ce discours dans un texte qui constitue le document relais à analyser ».

[5] La sémiologie (ou sémiotique) étudie précisément les liens entre systèmes sémiotiques, en se donnant pour objet « l'étude des signes et des processus interprétatifs », Ducrot et Schaeffer 1995, p. 179.

[6] Benveniste 1974 [1969], p. 43-66.

[7] « Il n'y a pas de "synonymie" entre systèmes sémiotiques ; on ne peut pas "dire la même chose" par la parole et par la musique, qui sont des systèmes à base différente ». *Ibid.* p. 53.

de notre analyse des notions de « document » et de « contenu », l'étude des documents non textuels, dans leur non-spécificité textuelle, doit être prise en compte.

Reste cependant que, essentiellement pour des commodités d'exposition, nous nous appuierons, en grande partie, sur des exemples d'indexation portant sur des documents textuels.

En outre ne seront pris en compte dans cette recherche que les textes de documents écrits en français ; cette restriction est liée aux types de théories linguistiques sollicitées qui travaillent essentiellement les phénomènes linguistiques du français. La dimension multilingue, nécessaire pour aborder l'indexation en toute généralité, pose des problèmes spécifiques qu'il ne nous est pas possible d'appréhender dans cette recherche.

III - Synthèse du chapitre et présentation du plan de l'étude

III.1 - Synthèse du chapitre

Dans ce chapitre, nous avons défini la problématique de notre recherche en la situant à deux niveaux : celui de l'objet d'étude et celui de la méthode d'analyse.

À chaque niveau, nous avons pu relever, dans les approches classiques de l'indexation, des impasses ou des contradictions, à partir desquelles nous avons formulé nos propres hypothèses. On rappelle ci-dessous l'essentiel de l'argumentation présentée :

- une approche de l'indexation définie en termes de recherche documentaire ne permet pas de conduire une étude de l'indexation sur un plan théorique, ceci pour deux raisons : d'une part, l'objet indexation n'y est pas autonomisable ; d'autre part, les présupposés des approches classiques de l'indexation conduisent à une impasse et à une contradiction : supposée être une technique, l'indexation devrait pouvoir faire l'objet d'une théorie, or c'est une théorie de la recherche documentaire qui fait l'objet de recherche ;

- en posant une distinction entre indexation et recherche documentaire, on fait l'hypothèse que l'indexation peut être appréhendée de façon autonome. En posant une distinction entre technique et pratique, on fait l'hypothèse que l'indexation, entendue comme une pratique, si elle ne peut faire l'objet d'une théorie, repose sur des représentations théoriques, qui peuvent se constituer en fondements théoriques, c'est-à-dire en fondements du point de vue d'une théorie. Là encore, deux hypothèses : d'une part, que l'indexation est une pratique empruntant à différentes théories ses modèles de fonctionnement ; d'autre part, que, parmi les représentations théoriques sous-jacentes à la pratique de l'indexation, les représentations linguistiques sont pertinentes pour constituer les fondements théoriques de l'indexation.

Dans le cadre de notre recherche, l'indexation, entendue comme une pratique reposant sur des fondements théoriques, est donc considérée de la façon suivante :

- l'indexation porte uniquement sur des documents mais sur tous les types de document, quelle que soit leur nature, textuelle ou non ; en revanche, on ne prend pas en compte dans cette recherche la notion d'indexation de requêtes ;

- l'indexation est à entendre comme un processus au sens « fort » : elle réalise des opérations de fabrication de ses objets, notamment les documents et les informations ; cette recherche devra montrer que la notion de processus au sens « faible » (notion de transfert) est insuffisante ;

- l'indexation se réalise par d'autres moyens que les « mots » : la notion de langage documentaire est à mettre à distance au profit d'autres procédés non plus lexicaux mais discursifs. Cette recherche devra montrer le caractère réducteur d'une approche de l'indexation qui ne considérerait que les « mots ».

Le cadre d'analyse retenu permet de donner corps à une telle approche de l'indexation ; en effet, par déplacements successifs de modèles, il donne la possibilité de :

- montrer comment, sur des points de conflits précis entre modèles formel et non formel de la langue, la description classique de l'indexation achoppe ;

- reconstruire, sur la base de ces lieux d'achoppement, de nouvelles représentations qui permettent de faire « voir » les fondements théoriques de l'indexation.

Cette étude est donc construite autour de deux principaux axes que nous présentons ci-après.

III.2 - Présentation du plan de la recherche

La recherche est menée sur quatre chapitres, regroupés en deux parties distinctes : une première partie essentiellement critique ; une seconde partie plus prospective. Chaque partie faisant l'objet d'une introduction spécifique, on en présente ici simplement les grandes lignes.

PREMIÈRE PARTIE : LES PROBLÈMES THÉORIQUES DE L'INDEXATION

Cette première partie emprunte son titre et son esprit à un ouvrage de Georges Mounin paru en 1963[1], dont la problématique nous paraît être très proche de la nôtre. Mounin y montre en effet en quoi l'activité professionnelle de la traduction constitue un « scandale » au regard des théories linguistiques[2] et pose les différentes pistes de recherche possibles une fois ce constat posé[3] :
(i) on peut remettre en cause le bien fondé d'une pratique ;

---

[1] Mounin 1963 : *Les Problèmes théoriques de la traduction.*
[2] *Ibid.*, p. 8 : « On pourrait presque dire que l'existence de la traduction constitue le scandale de la linguistique contemporaine ».
[3] *Ibid.*, p. 8-9.

(ii) on peut remettre en cause le bien fondé des théories linguistiques ;
(iii) on peut chercher à comprendre les raisons d'une cohabitation entre impossibilité théorique et possibilité empirique.

Si, tout comme Mounin, nous considérons que seule la voie (iii) est constructive, il importe de commencer par mettre au jour les incompatibilités entre modèle d'utilisation de la langue en indexation et modèle de fonctionnement de la langue en linguistique.

Pour mettre au jour ces incompatibilités, on s'intéresse aux lieux de « tension » entre ces deux modèles (les lieux où les modèles se « disputent les mêmes objets » dans les termes de Berrendonner et Reichler-Béguelin) ; nous en avons identifié deux, que nous traitons respectivement dans les chapitres II et III qui constituent cette première partie : le lexique (chapitre II) et la référence (chapitre III).

DEUXIÈME PARTIE : CONTRIBUTION AUX FONDEMENTS THÉORIQUES DE L'INDEXATION

La première partie fait apparaître que le modèle de fonctionnement de la langue sous-jacent à la pratique de l'indexation s'apparente, sur les questions du lexique et de la référence, au modèle du sens commun, contre lequel s'est notamment construite la linguistique. On peut choisir d'en rester là, mais on ne peut alors dégager des fondements théoriques de l'indexation.

On peut aussi envisager, dans la perspective de constituer les fondements de l'indexation du point de vue de la théorie linguistique, de substituer à ce modèle de sens commun un modèle scientifique : c'est ainsi que l'on propose, dans cette seconde partie, un nouveau modèle d'utilisation de la langue qui s'inscrit, lui, dans le cadre d'un modèle linguistique. On est amené à redéfinir l'indexation sous ses deux principaux aspects, processus et résultat :

- appréhendée sous l'angle du processus, l'indexation est analysée dans une perspective discursive : cette étude fait l'objet du chapitre IV ;

- appréhendée sous l'angle du résultat, l'indexation est réinterrogée dans ses « formes » : formes logique et linguistique du descripteur, traitées dans le chapitre V.

Cette deuxième partie propose des éléments pour constituer les fondements théoriques de l'indexation, éléments qui, établis dans le cadre d'un modèle d'utilisation de la langue, relèvent d'emprunts et de transformations de concepts issus de la linguistique. Sur ce point, le modèle d'utilisation de la langue proposé en deuxième partie, s'il ne sort pas à proprement parler du « bricolage », permet de souligner ce que l'indexation gagne à exploiter l'« hétérogénéité » des textes et les « ambiguïtés » de la langue.

Nous espérons ainsi montrer en quoi le modèle d'analyse adopté peut se révéler fécond. Il permet de penser plusieurs types d'approche de l'indexation : l'indexation classique qui se greffe, faute de mieux et par commodité, sur le modèle du sens commun, mais qui du même coup se heurte, sans pouvoir les contourner, aux propriétés de la langue ; une approche discursive de l'indexation qui cherche à prendre en compte, à tirer partie de la langue, ambiguë, variée, hétérogène.

Première partie

Les problèmes théoriques de l'indexation

En reformulant la démarche de Mounin [1963] dans le cadre de notre domaine (l'indexation) et de notre problématique (l'étude des fondements théoriques), nous appellerons « problèmes théoriques » de l'indexation les cas d'incompatibilité entre modèle d'utilisation et modèle de fonctionnement de la langue : c'est-à-dire les cas où les descriptions linguistiques proposées, implicitement ou explicitement, par les professionnels de l'indexation ne sont pas avérées comme des descriptions de faits de langue, telles que la linguistique contemporaine a pu les conduire.

Précisons, avant d'entamer une telle étude :

- les lieux d'inscription du modèle d'utilisation de la langue en indexation (A) ;
- l'enjeu d'une étude des problèmes théoriques de l'indexation (B).

A - Le modèle d'utilisation de la langue en indexation

Le modèle d'utilisation de la langue en indexation, ou encore la description linguistique (implicite et explicite) des faits d'indexation, se capte difficilement au niveau des pratiques d'indexation[1] elles-mêmes. C'est essentiellement au niveau des discours sur ces pratiques qu'un tel modèle se révèle. Ce sont donc ces discours que nous étudierons.

Ces discours de la pratique sur elle-même, que nous appelons discours classiques ou discours normatifs, peuvent être de nature différente mais présentent deux caractéristiques communes :

---

[1] Le terme « pratique » est ici à comprendre dans son acception courante comme « manière concrète d'exercer une activité », *Le Robert 1* 1993.

- ces discours s'adressent aux professionnels dans le but de constituer un référentiel commun, nécessaire notamment à l'harmonisation et à l'enseignement des pratiques. C'est particulièrement le cas des textes normatifs[1] et des textes didactiques[2] ; c'est aussi le cas des traités[3] qui, parfois sous le nom de « théories de l'indexation », proposent le plus souvent une approche synthétique et globalisante des règles d'indexation formulées d'un point de vue normatif ;

- ces discours ont pour point de départ et pour point d'arrivée les pratiques d'indexation elles-mêmes. Le caractère général des descriptions qu'ils donnent tient plus d'une abstraction des contextes particuliers d'application que d'une véritable « formalisation ». L'indexation reste décrite dans son propre domaine, du point de vue de l'indexation elle-même : elle n'est pas objectivée[4] ; en ce sens, le référentiel constitué par ces discours reste proprement documentaire.

B - Enjeu d'une étude des problèmes théoriques de l'indexation

Le discours de la pratique sur elle-même présente la particularité de ne s'inscrire dans aucun référentiel théorique explicite. Ce n'est pas pour autant qu'il ne dispose d'aucun arrière-plan théorique implicite. C'est là une particularité des pratiques professionnelles : Mounin [1963], sur les aspects de la traduction, et Corbin [1987], dans son étude des pratiques lexicographiques, relèvent de la même façon que, pour peu que l'on puisse rapporter une pratique à une théorie, cette théorie s'apparente toujours à une théorie du sens commun. L'inscription d'une pratique dans un arrière-plan théorique apparaît inévitable, mais cette inscription ne relève pas toujours d'un choix explicite. Quand il reste implicite, le choix d'un arrière-plan théorique résiste mal à l'évidence des modèles du sens commun.

C'est ainsi que les discours de la pratique d'indexation sur elle-même empruntent, entre autres, les formes de ce que nous avons appelé la *doxa* linguistique. Cette *doxa* linguistique présente la particularité de ne pas permettre de distinguer le niveau de la langue et le niveau de l'utilisation de la langue. De là les problèmes théoriques de l'indexation, les problèmes d'indistinction et de chevauchement entre faits de langue et faits d'indexation. Cette indistinction de niveaux permet difficilement de définir en propre l'indexation : l'utilisation particulière de langue qui s'y fait ne peut apparaître en tant que telle.

L'enjeu d'une étude des problèmes théoriques de l'indexation est de permettre de dégager la spécificité de l'indexation. On voudrait pouvoir montrer que l'indexation

---

[1] Dans le domaine français, les textes normatifs de base sont : la norme Z 47-102 (Principes généraux pour l'indexation des documents) 1978 ; la norme Z 47-100 (Règles d'établissement des thésaurus monolingues) 1981 ; la norme Z 47-200 (Liste d'autorité de matières) 1985.

[2] Parmi les productions récentes de manuels d'indexation, Chaumier 1996.

[3] Pour les traités théoriques les plus connus, Fugmann 1993 et Lancaster 1991.

[4] On veut dire par là que l'indexation reste, dans le discours classique, un objet empirique ; elle n'est pas constituée comme objet scientifique.

met en œuvre une utilisation professionnelle de la langue et qu'en cela elle est contrainte de suivre les propriétés de la langue elle-même : nous rejoignons là encore les ambitions de Mounin[1].

Cette partie s'attache à l'étude de deux problèmes théoriques de l'indexation : le chapitre II s'intéresse à la question du lexique, le chapitre III à celle de la référence[2]. Les deux chapitres suivent une même démarche – analyse, déplacement, proposition : analyse du discours classique, déplacement de l'arrière-plan théorique et proposition d'un nouveau cadre de formulation des faits d'indexation.

La construction de ce nouveau cadre reste, dans ces deux chapitres, partielle : la seconde partie de cette étude sera spécifiquement vouée à la mise au point d'un modèle d'utilisation de la langue en indexation.

---

[1] Mounin 1963, p. 16-17 : « La traduction [...] comporte certainement des aspects franchement non linguistiques, extra-linguistiques. Mais, toute opération de traduction – Fédorov a raison – comporte, à la base une série d'analyses et d'opérations qui relèvent spécifiquement de la linguistique, et que la science linguistique appliquée correctement peut éclairer plus et mieux que n'importe quel empirisme artisanal. On peut, si l'on y tient, dire que, comme la médecine, la traduction reste un art – mais un art fondé sur une science. Les problèmes théoriques posés par la légitimité ou l'illégitimité de l'opération traduisante, et par sa possibilité ou son impossibilité, ne peuvent être éclairés en premier lieu que dans le cadre de la science linguistique ».

[2] Si lexique et référence constituent des problématiques intrinsèquement liées en linguistique, elles sont ici étudiées de façon séparée pour faire apparaître clairement leur rôle respectif en indexation.

CHAPITRE II
LA QUESTION DU LEXIQUE EN INDEXATION

La question du lexique[1] en indexation constitue, de façon typique, pourrait-on dire, une zone de tension entre modèle d'utilisation et modèle de fonctionnement de la langue.

Les deux modèles se disputent en effet les « mêmes objets » : les « mots[2] » – descripteurs* dans le premier modèle, unités lexicales* dans le second – appréhendés selon des points de vue, des modes de saisie radicalement différents :

- le discours classique refuse la notion de lexique, au nom de son inorganisation et de son ambiguïté[3]. Il y substitue la notion de langage documentaire, dont la visée pragmatique est censée corriger les imperfections relevées. Le mode de saisie de ses unités se veut déconnecté des notions intuitive de « mot » et linguistique d'« unité lexicale[4] » ;

- l'approche linguistique refuse, elle aussi, la notion intuitive de lexique, notamment parce qu'y est privilégié un point de vue « externe » sur le lexique (vu comme une instance d'enregistrement d'expressions linguistiques « données ») au détriment d'un point de vue « interne[5] ». Considéré d'un point

---

[1] Dans cette première occurrence, le lexique est entendu dans son acception « intuitive » comme l'ensemble des mots d'une langue.
[2] On emploie ici encore, pour le début de ce chapitre, une notion intuitive, celle de « mot », sans chercher à la définir. On emploiera également le terme de « forme lexicale » pour renvoyer à cette même notion intuitive de « mot ».
[3] Par exemple, la norme AFNOR Z 47-100 (1981), p. 185 : « Un thésaurus ne doit être confondu ni avec un lexique, ni avec un index, ni avec un dictionnaire. [...] Le vocabulaire constituant le thésaurus doit être *non ambigu* pour que les mêmes termes de ce vocabulaire identifient systématiquement les mêmes concepts ; *structuré* pour assurer une meilleure définition de chaque terme et pour permettre des recherches à différents degrés, de généralité ou de spécificité ». (C'est nous qui soulignons).
[4] Chaumier 1978, p. 29-30 : la « signification » des descripteurs est fixée par la structure des thésaurus notamment grâce aux relations hiérarchiques qui « ne sont pas nécessairement basées sur des critères sémantiques mais aussi sur des considérations pragmatiques résultant des besoins de la recherche documentaire ».
[5] Marandin 1992a [présentation], p. 6.

de vue interne, le lexique constitue un artefact[1], susceptible de faire l'objet d'une théorie, dans laquelle les unités lexicales peuvent trouver des formes de régularité et d'organisation[2].

La notion intuitive de lexique est donc critiquée dans les deux approches mais selon des points de vue différents : en raison de contraintes d'utilisation dans le modèle de l'indexation, en raison de critères scientifiques dans le modèle linguistique. Cependant, dans les deux cas, on substitue pareillement, à la notion intuitive de lexique, la notion de modèle : modèle de fonctionnement en linguistique, modèle d'utilisation en indexation, mais modèle d'utilisation qui, dans le discours classique, se donne pour ou qui dit se passer d'un modèle de fonctionnement.

Or, en étudiant les différentes fonctions attribuées au descripteur, on constate que, loin de prendre ses distances avec un modèle de fonctionnement de la langue, le modèle d'utilisation du lexique en indexation repose sur les conceptions les plus traditionnelles et les plus courantes du lexique. On peut en effet relever les formes d'une *doxa* linguistique[3] qui constituent implicitement l'arrière-plan théorique du modèle d'utilisation du lexique en indexation. Cet arrière-plan théorique révèle, sur un plan pratique, des limites et des insuffisances, qui sont alors imputées, par les professionnels, à la langue elle-même ; or c'est à une certaine vision du lexique qu'elles devraient plutôt être rapportées. L'arrière-plan théorique implicitement présent dans le discours classique présente en outre l'inconvénient de ne pouvoir rendre compte des propriétés des unités lexicales réellement en jeu dans l'indexation. Comme nous le montrerons, la *doxa* linguistique joue, dans la description classique de l'indexation, un rôle fortement opacifiant.

Pour tenter de démêler ces différents types et niveaux de représentations linguistiques en jeu dans cet aspect de l'utilisation des « mots » en indexation, nous procéderons de la façon suivante :

- à partir du discours de la norme[4], qui constitue le cœur des approches classiques, nous étudierons le modèle du lexique en indexation sous ses deux aspects : le modèle d'utilisation du lexique (les différentes fonctions du descripteur) et le modèle de fonctionnement sous-jacent (la *doxa* linguistique) ;

- nous examinerons ensuite les fonctions attribuées au descripteur à la lumière d'un autre modèle de fonctionnement, lui explicitement linguistique, des unités lexicales : nous pourrons distinguer, parmi ces différentes fonctions, celles qui relèvent d'effets propres à une utilisation et celles qui relèvent de principes propres à la langue ;

- suite à ce déplacement de modèles de fonctionnement du lexique, nous reformulerons le modèle d'utilisation du lexique en indexation. La distinction des faits et des effets nous conduira à mettre à distance le rôle des « mots » en

---

[1] Rastier 1994, p. 28.
[2] Marandin 1992a [présentation], p. 6 : « La question d'une théorie autonome du lexique se pose dès lors que l'on suppute que les caractères d'inorganisation, d'irrégularité du lexique, d'idiosyncrasie des entrées sont davantage liés au point de vue adopté qu'à l'objet que l'on discerne sous la construction théorique, à savoir le matériel lexical propre à une langue ».
[3] Rappelons que, dans cette recherche, la notion de « *doxa* linguistique » renvoie à l'ensemble des représentations que tout sujet parlant se forge sur la langue et le langage.
[4] Norme AFNOR Z 47-102 (Principes généraux pour l'indexation des documents) 1978.

indexation et ce à deux niveaux : au niveau de la description du processus de l'indexation, dont on proposera une nouvelle formulation, et au niveau de la définition des descripteurs, dont on réévaluera les fonctions.

I - Le modèle du lexique en indexation

Si, comme nous l'avons précédemment relevé, il n'est pas question, à proprement parler, de lexique dans le discours classique, il y est en tout cas question des « mots », et plus précisément du rôle que ceux-ci sont censés jouer en tant que descripteurs[1]. Il nous paraît important de ne pas masquer, ne serait-ce qu'à travers la notion confuse de « mot » dans un premier temps, la dimension lexicale des descripteurs puisque, malgré l'idéal qu'elle se donne[2], l'indexation continue d'emprunter à la langue son matériel lexical, alors même que certains chercheurs ont pu proposer des langages documentaires « véritablement » artificiels[3].

C'est en fonction de cette approche du descripteur que l'on dégagera les deux niveaux du modèle du lexique en indexation : niveau de l'utilisation et niveau des représentations linguistiques sous-jacentes.

---

[1] Notre approche nous conduit à considérer la notion de descripteur dans une acception plus large que celle que lui accorde la norme (voir glossaire). Le descripteur sera à entendre dans cette recherche, sauf indication contraire, comme toute forme lexicale utilisée pour indexer, que cette forme lexicale fasse partie ou pas d'un langage documentaire. En effet, la distinction entre indexation libre (indexation par choix libre de termes) et indexation contrôlée (indexation par choix dans une liste close) que sous-tend la définition normative du descripteur ne peut être prise en compte à ce stade de l'étude : elle implique en effet une distinction que nous discutons, qui consiste à poser une différence de « nature » entre mots de la langue et descripteurs là où nous proposons de voir une distinction d'« utilisation ».

[2] C'est chez un linguiste que l'on trouve, selon nous, la meilleure expression de cet idéal : « Ce qui est souvent admis comme résultat idéal pour un automate dans ce domaine [documentaire] c'est un texte en langue naturelle vidé des supports linguistiques des chaînes et présentant en leur lieu et place des identifieurs symboliques, des indices. Soit à peu près : remplacer, en utilisant des nombres ou des lettres, les expressions d'un texte par le symbole de ce qui est désigné », Corblin 1987, p. 15.

[3] C'est toute la réflexion de Gardin et de son équipe menée dans les années 1950 sur le langage « mécanographique » dans le cadre du système SYNTHOL (*SYNTagmatic Organization Language*). Leur recherche portait sur la façon d'« obtenir une analyse sémantique des objets qui soit indépendante de leurs noms dans les langues naturelles » ; ils proposaient, pour ce faire, la constitution de véritables codes, fondés exclusivement sur des symboles (lettres ou chiffres), à l'exclusion de toute forme lexicale. On peut se reporter à Bely *et al.* 1970, Gardin 1967 et Gardin 1974 ; dans Gardin 1991, on trouve une synthèse et une évaluation d'une partie des travaux menés dans le cadre du système SYNTHOL.

I.1 - Modèle d'utilisation du lexique en indexation

En reprenant l'esprit de la norme[1], on peut dire que l'indexation d'un document* se traduit par l'attribution à un document d'un ou de plusieurs mots, ces mots entretenant avec le document un rapport de deux types :

- un rapport de représentation : l'indexation fournit la représentation du contenu d'un document. Cet aspect correspond, dans le texte de la norme, à la notion d'indication de la « teneur » d'un document ;

- un rapport de catégorisation : l'indexation indique l'appartenance d'un document à un (ou plusieurs) ensemble(s) de documents jugés semblables. Cet aspect correspond, dans le texte de la norme, à la notion d'accès non à un document mais « aux informations contenues dans un fonds documentaire ».

À partir de cette première paraphrase, on dira que le descripteur se voit attribué, dans la norme, deux fonctions : il est à la fois une unité de représentation du contenu d'un document et une unité d'accès à l'information d'un ensemble documentaire.

L'indexation pose donc le descripteur dans un rapport à deux niveaux distincts, sans que l'on sache comment s'établit le passage de l'un à l'autre : pensé dans ses rapports avec un document, le descripteur s'approche en termes de contenu ; pensé dans ses rapports avec un fonds documentaire, il est décrit en termes d'accès.

Pour essayer de préciser ces deux fonctions du descripteur et le lien éventuel qu'elles entretiennent, nous commencerons par les examiner chacune séparément.

I.1.1 - LE DESCRIPTEUR EN TANT QU'UNITÉ DE REPRÉSENTATION DU CONTENU D'UN DOCUMENT

On peut comprendre ce que le discours normatif entend par les notions de « contenu » et de « représentation » en examinant la description du processus de l'indexation qu'il donne.

L'indexation y est décrite comme un processus réalisé en deux phases :

- une phase d'analyse de contenu, dite phase d'analyse conceptuelle où s'effectue « la reconnaissance des concepts contenant l'information dans les documents à indexer » ;

- une phase de représentation dite phase de traduction correspondant à la « représentation de ces concepts dans le langage documentaire ».

Cette appréhension des notions de « contenu » et de « représentation » détermine fortement les caractéristiques dont doit être pourvu le descripteur : nous les précisons ci-après.

---

[1] Voir les deux fonctions de l'indexation : « permettre la recherche efficace des informations contenues dans un fonds documentaire » et « indiquer brièvement, sous forme concise, la teneur du document ». Norme AFNOR Z 47-102 1978, p. 225.

A - La notion de « contenu » : le descripteur comme expression linguistique de concept

Il apparaît que la notion de contenu est, de façon générale en indexation, de nature conceptuelle : qu'il s'agisse de caractériser le contenu d'un document ou le contenu d'un descripteur, c'est toujours à la notion de « concept » que l'on se réfère. S'il est cependant plutôt question de « notion » pour qualifier le contenu d'un descripteur[1] et de « concept » pour qualifier celui d'un texte, les deux termes sont en fait unifiés par le biais d'une même définition[2].

Le descripteur, comme élément d'un langage documentaire ou comme élément de description d'un document est, dans tous les cas, l'expression linguistique d'un concept. Cette caractéristique repose sur les présupposés suivants :

- les concepts sont donnés comme préexistants, c'est-à-dire comme existant préalablement à toute formulation linguistique (le descripteur est un terme qui « renvoie » à un concept) comme à toute analyse documentaire (les concepts sont « reconnus » et « extraits » d'un document) ;

- les concepts sont donnés comme stables : les concepts d'un document sont les mêmes que ceux des descripteurs qui sont les mêmes que ceux des requêtes documentaires[3] ; autrement dit, la formulation linguistique d'un concept n'obère pas son appréhension[4] ;

- les concepts de l'indexation sont des concepts « simples » qui doivent correspondre, idéalement, à des mots « simples ». L'assimilation concept/mot se joue en indexation jusque dans la détermination des formes du descripteur : la forme linguistique du descripteur doit être « décomposée » en fonction d'une « décomposition » conceptuelle[5].

Pour être une unité de représentation du contenu d'un document, le descripteur doit donc fonctionner comme l'expression linguistique d'un concept préexistant, stable et simple. La norme ne précise pas si c'est cette expression linguistique du concept

---

[1] Voir la définition normative du langage documentaire : « langage artificiel constitué de *représentations de notions* et de relations entre ces notions et destiné, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs ». *Vocabulaire de la documentation*, 1987 (c'est nous qui soulignons).

[2] La norme AFNOR [Z 47-102 (1978), p. 231] assimile « notion » et « concept » par une définition commune : « toute unité de pensée ».

[3] La croyance en une stabilité du concept se situe au cœur du projet des langages documentaires, cf. norme Z 47-100 1981, p. 3 : « Le thésaurus permet donc de traduire en termes d'indexation ou en termes de recherche *tout concept devant entrer ou sortir d'un système documentaire donné* ». C'est nous qui soulignons.

[4] Cette idée apparaît également de façon claire chez Ranganathan [1967], qui distingue, lui, trois phases dans l'indexation (« Idea Plane », « Verbal Plane », « Notational Plane ») : « subject analysis takes place in an Idea Plane which *words we use to express our ideas is not important* ». C'est nous qui soulignons.

[5] Cf. norme 47-100 1981, p. 186 : dans le cas d'« expressions complexes », il faut décomposer de façon parallèle mots et concepts. « Il existe deux manières d'analyser les termes complexes pour y reconnaître les notions simples :
- l'analyse sémantique, exemple : oxycoupage = découpage+oxygène ;
- l'analyse morphologique (syntaxique), exemple : psychologie des animaux = psychologie+animal ».

ou le concept lui-même qui constitue une information, une partie de l'information d'un document.

B - La notion de « représentation » : le descripteur comme expression linguistique

La notion de représentation relève, dans le discours classique, de deux plans différents :

- sur un plan lexical, il s'agit de représenter un concept, de « mettre en mots » les concepts stables et simples préexistants. Cette opération s'effectue sur le mode de la « traduction », elle-même de deux types : une « traduction » de type conceptuel (concept –> mot), une « traduction » de type interlingual (mot_1 –> mot_2[1]) ;

- sur un plan textuel, il s'agit de représenter un texte, c'est-à-dire de réduire le texte du document à indexer. La réduction porte elle aussi sur deux plans[2] : celui de la « forme » (l'indexation, c'est le texte original en plus petit) et celui du « fonds » (l'indexation ne dit pas tout du texte, seulement ce qui est le plus essentiel[3]).

Ces deux aspects de la représentation en indexation reviennent à attribuer des caractéristiques particulières au descripteur :

- le descripteur doit fonctionner comme une unité stable (mais on ne sait pas s'il s'agit de la même stabilité que celle reconnue par ailleurs au concept) : il doit en effet maintenir le « concept » d'origine à travers les différentes formes linguistiques qui peuvent servir à l'exprimer ;

- le descripteur doit fonctionner comme un condensateur textuel : être une expression du texte à la fois réduite du point de vue de la forme et synthétique du point de vue de son « sens ».

---

[1] Où « mot_1 » est le premier terme venant à l'esprit de l'indexeur pour exprimer le concept et « mot_2 » le mot du langage documentaire, le descripteur, s'approchant au mieux du « mot_1 ».

[2] Ces deux aspects de la réduction ne sont pas toujours distingués. On ne sait pas toujours, chez certains auteurs, si l'indexation opère une condensation de tout le texte (avec les pertes « informationnelles » liées au principe de la condensation) ou une focalisation sur certaines parties du texte. On trouve, en effet, des formulations diverses : l'indexation relève chez les uns d'une représentation exhaustive du texte d'origine (Dewèse 1993, p. 165 : « moyen d'expression réduite du texte intégral » ; Bertrand-Gastaldi 1986, p. 4 : « opération qui vise à substituer au texte de départ un texte d'arrivée, plus court et mieux organisé, plus économique à mémoriser et à manipuler »). Chez d'autres auteurs, l'indexation relève d'une représentation sélective du texte (Chaumier 1988 : « expression plus ou moins condensée des caractéristiques d'un document » ; de même la norme Z 47-102, qui préconise une « sélection des concepts », voir note 3 suivante).

[3] Outre que l'indexation n'explicite pas toujours ce qu'elle choisit de dire d'un texte, les critères de sélection sur lesquels elle repose paraissent difficiles à formaliser : tout dépend de la « fonction », c'est-à-dire de la fameuse finalité donnée à l'indexation. Sur ce point, voir, par exemple, la norme Z 47-102, 1978, p. 227 : « Tous les concepts identifiés par l'indexeur comme représentatifs de l'information contenue dans le document ne sont pas nécessairement retenus pour l'indexation. En effet, l'indexation doit essentiellement être adaptée à sa *fonction* propre dans le système où elle est utilisée ». C'est nous qui soulignons.

Pour pouvoir fonctionner comme une unité de représentation du contenu d'un document, un « mot » donc doit relever simultanément :
- *du concept : il doit être une expression conceptuelle, pouvoir exprimer un concept ;*
- *du symbole : il doit être une expression stable, pouvoir exprimer toujours le même concept ;*
- *du texte : il doit être une expression textuelle, pouvoir exprimer un texte ou certaines de ses parties.*

En plus de ces trois caractéristiques liées à la première de ses fonctions s'en ajoutent d'autres, liées, elles, à la fonction du descripteur comme accès à un ensemble documentaire.

I.1.2 - LE DESCRIPTEUR EN TANT QU'ACCÈS À UN ENSEMBLE DOCUMENTAIRE

Le discours classique détaille fort peu cette fonction de l'indexation (permettre la recherche d'information) et cet aspect du descripteur (accès à l'information d'un ensemble documentaire).

On retrouve là la dissolution de l'objet « indexation » propre aux approches classiques : insensiblement le discours sur l'indexation se tait et cède la place à celui sur la recherche documentaire.

Cependant, il y a implicitement, dans le discours classique, l'idée que l'accès à l'information d'un fonds documentaire se fait sur la base des unités de représentation de contenu de chacun des documents qui le constituent[1].

Il revient donc bien à un modèle de l'indexation (et non à un modèle de la recherche documentaire) d'expliquer comment peut se réaliser ce double saut :
- saut du document à une pluralité de documents ;
- saut du « contenu » à l'« information ».

Il importe donc de déterminer les caractéristiques du descripteur en fonction de ces deux rôles.

Le modèle d'utilisation du lexique en indexation ne dit rien de la fonction du descripteur comme accès à l'« information » contenue dans plusieurs documents différents. Dans ce paragraphe, nous laissons donc ce point de côté ; nous le reprendrons dans le cadre d'un autre modèle. En revanche, il nous fournit quelques éléments sur le fonctionnement du descripteur comme accès à un document d'une part et comme accès à plusieurs documents d'autre part.

A - L'accès à un document : le descripteur comme une expression linguistique autonome

Le discours classique ne relève ni ne commente le fait que les descripteurs retenus pour indexer un document puissent être lus, interprétés et utilisés, seuls, détachés du document et/ou du langage documentaire d'où ils proviennent. Il y a là une

---

[1] On trouve chez Menon une formulation explicite de ce lien entre représentation du contenu et accès à l'information : « Nous nous intéresserons surtout ici aux unités susceptibles de permettre l'accès à l'information par des éléments tirés de son contenu thématique », Menon 1988, p. 146.

évidence que l'ensemble des mots issus de l'indexation doit exprimer quelque chose pour un utilisateur (mais quoi ? le contenu d'un document ? de l'information ? un sens ?).

De cette évidence, qui ne va pas de soi[1], on peut néanmoins dégager une caractéristique dont doit être doté le descripteur : il doit révéler une certaine autonomie, lui permettant de fonctionner seul. Nous aurons à déterminer le type d'autonomie dont le descripteur doit être pourvu : conceptuelle ? référentielle ? sémantique ?

B - L'accès à plusieurs documents : le descripteur comme relais textuel

La norme suggère que le descripteur joue un rôle non seulement par rapport à un document mais aussi par rapport à plusieurs documents : il doit permettre, au sein d'un fonds documentaire, d'établir des liens, d'effectuer des rapprochements, de constituer des ensembles. Un même descripteur doit donc pouvoir être affecté à des documents différents, soit parce que les « concepts qui s'y trouvent » sont jugés semblables, soit parce que les « concepts qui s'y trouvent » sont posés comme équivalents au concept que « représente » le descripteur.

Il y a donc dans la notion de descripteur celle de classe d'équivalence : est-ce une classe d'équivalence lexicale (de « mots ») ? Est-ce une classe d'équivalence documentaire (de « choses ») ?

La norme n'est pas très claire sur ce point. Si elle dit assez nettement qu'il ne s'agit pas de classe d'équivalence lexicale, elle ne dit pas explicitement qu'il s'agit de classe d'équivalence documentaire. En effet, le descripteur y est défini comme étant un « terme préférentiel », c'est-à-dire un terme retenu de préférence à d'autres au sein d'un ensemble de termes jugés équivalents. Dans la norme, le jugement d'équivalence ne s'établit pas en fonction du « sens » des termes[2] mais en fonction de leur emploi supposé[3], ce qui renvoie à nouveau le problème dans le camp de l'utilisateur et de la recherche documentaire[4].

Même si elles sont peu détaillées dans le discours classique, la présence de la notion de classes d'équivalence – plutôt de type documentaire que lexical –, et celle, corollaire, de synonymie, signalent qu'un descripteur doit être à même d'indiquer un

---

[1] On peut toujours trouver un « sens » à une série de mots : c'est ce que David et Plante [1990a] ont proposé d'appeler le « syndrome de la signification garantie ». En revanche, on ne peut jamais être sûr que ce sens corresponde à celui du texte dont ces mots sont extraits ou pour lequel des mots ont été choisis.

[2] Chaumier 1978, p. 33 : « La notion de synonymie est utilisée de façon extensive dans les thésaurus sous la forme de la synonymie documentaire afin de regrouper sous un seul descripteur plusieurs termes considérés comme voisins bien que de signification sémantique différente ».

[3] Norme Z 47-100 (1981), p. 189 : « L'établissement des relations d'équivalence doit être gouverné par le principe simple suivant : lorsque tout document indexé à l'aide du terme A doit être pris en considération pour toute demande indexée à l'aide du mot B (et réciproquement), les termes A et B sont synonymes documentaires (soit vrais synonymes soit quasi-synonymes). Il suffit alors de choisir l'un des termes A ou B comme descripteur, l'autre étant non-descripteur ».

[4] *Ibid.*, p. 190 : « Le choix des termes à utiliser comme descripteurs à l'intérieur des classes d'équivalence doit obéir à des critères à établir à partir des besoins de la majorité des utilisateurs potentiels ».

jugement de ressemblance entre documents : on dira qu'en cela le descripteur doit jouer un rôle de relais entre textes, de relais textuel.

I.1.3 - CARACTÉRISTIQUES DU DESCRIPTEUR : RÉCAPITULATIF

Sous forme de récapitulatif, on rappelle comment les deux fonctions attribuées à l'indexation dans le texte de la norme déterminent celles du descripteur et comment ces fonctions du descripteur peuvent se laisser traduire en caractéristiques dont il doit être doté[1] :

- **fonction 1 de l'indexation** : « indiquer brièvement, sous forme concise, la teneur du document »
  fonction 1 du descripteur : le descripteur comme unité de représentation du contenu d'un document :
-> caractéristique 1 : le descripteur comme expression linguistique d'un concept préexistant, stable, unitaire (« simple ») ;
-> caractéristique 2 : le descripteur comme expression linguistique stable ;
-> caractéristique 3 : le descripteur comme condensateur textuel.

- **fonction 2 de l'indexation** : « permettre la recherche efficace des informations contenues dans un fonds documentaire »
  fonction 2 du descripteur : le descripteur comme accès à un fonds documentaire :
-> caractéristique 4 : le descripteur comme expression linguistique autonome ;
-> caractéristique 5 : le descripteur comme relais textuel.

Dans ce paragraphe, on a tenté de montrer comment s'organisait le modèle d'utilisation des mots en indexation. À partir des fonctions de l'indexation se déduisent celles du descripteur. Ces fonctions font l'objet d'une description, qui, pour ne pas être toujours complète, permet néanmoins à l'indexation de se constituer en pratique[2]. Cette description revient à attribuer, de façon explicite et implicite, des caractéristiques au descripteur : c'est lors de ce passage des fonctions aux caractéristiques du descripteur que le discours classique emprunte à un modèle courant du lexique ses présupposés, comme nous le verrons ci-après.

Il semble en effet que, pour être utilisés de la façon dont l'indique la norme, les descripteurs n'ont pas nécessairement besoin d'être pourvus des caractéristiques qui leur sont attribuées explicitement et implicitement : il y a, dans l'approche du descripteur dans le discours normatif, l'adoption implicite de représentations linguistiques qui, pour être communes à l'ensemble des sujets parlants, ne s'en révèlent pas moins à la fois inadaptées et limitées pour décrire les faits d'indexation.

I.2 - Modèle de fonctionnement implicite du lexique en indexation

Le modèle d'utilisation du lexique en indexation, précédemment exposé, repose sur l'articulation de deux types de représentation linguistique des plus courants, représentations que les sciences du langage ont pu remettre en question :

---

[1] La numérotation adoptée ci-dessous ne traduit aucune idée d'ordre (dans le processus de l'indexation par exemple) ; elle est utilisée pour faciliter les renvois ultérieurs.
[2] Dans l'acception courante du terme, « manière concrète d'exercer une activité ».

- le premier type concerne la nature même de la langue, appréhendée sous le seul angle de ses formes lexicales : c'est ce que l'on nommera le modèle lexicaliste ;
- le second type, corollaire du premier, concerne la fonction du langage comme instrument de communication : ce présupposé relève de ce qu'il est convenu d'appeler un modèle objectiviste du langage[1].

I.2.1 - UN MODÈLE « LEXICALISTE » DE LA LANGUE

A - Approche du modèle lexicaliste

On aura pu remarquer, dans le précédent paragraphe, que l'ensemble du processus de l'indexation passe par les mots et ne met en jeu que des mots : les concepts d'un texte sont des mots, les concepts des descripteurs sont des mots, les ressemblances entre textes se font sur la base de mots, etc.

Les mots en indexation assimilent donc, par le biais des « descripteurs », des objets de nature différente (notamment les objets cognitifs et les objets textuels[2]).

On qualifiera cette approche de « lexicaliste » pour signifier qu'elle ne pense la langue que sous un seul rapport : celui des formes lexicales.

B - Critiques du modèle lexicaliste

L'un des enjeux de la linguistique contemporaine a été précisément de montrer que la langue était constituée de plusieurs dimensions[3] et que les formes lexicales ne pouvaient constituer l'accès unique et exhaustif au « sens » ou encore au « contenu » des textes. L'une des premières tâches de la linguistique a donc consisté à démonter la vision de la langue comme nomenclature, comme répertoire. C'est Saussure qui le premier a mené la critique[4], relayée depuis par des chercheurs de différents courants, structuralistes et non structuralistes. Plusieurs aspects du modèle de la langue comme nomenclature ont été diversement discutés dans le champ linguistique, notamment celui du sens des mots.

En effet, le modèle de la langue-répertoire conduit à envisager la dénomination* sous la forme d'un baptême, selon une procédure canonique dont on peut voir, comme le propose Mounin[5], la première expression dans la Bible : « Et Dieu nomma la lumière Jour, et les ténèbres Nuit [...] Et Dieu nomma l'étendue Cieux », etc. Dans ce modèle, les mots n'ont pas de signification qui leur est propre : ils sont perçus comme des symboles, des conventions interchangeables. Or, quelle que soit leur approche, les linguistiques contemporaines accordent toutes une place à la signification lexicale, comprise comme l'une des dimensions susceptibles de

---

[1] Dubois 1995.
[2] Comme le fait remarquer Dubois 1995, p. 88-89.
[3] Par exemple, Milner 1989.
[4] Sur la base de départ suivante : « Pour certaines personnes, la langue, ramenée à son principe essentiel, est une nomenclature, c'est-à-dire une liste de termes correspondant à autant de choses [...]. Cette conception suppose des idées toutes faites préexistant aux mots », Saussure 1973, p. 97.
[5] Mounin 1963, p. 25-26.

donner une « individualité » singulière aux « mots », permettant de les distinguer les uns des autres[1].

C - Marques du modèle lexicaliste en indexation

Pour discutée qu'elle puisse être dans le champ linguistique, cette approche de la langue comme nomenclature et de la dénomination comme baptême n'en constitue pas moins le modèle de représentation courant de la langue, un type de représentation linguistique standard, pourrait-on dire[2]. Que le modèle de l'indexation l'adopte n'est donc pas en soi étonnant et n'aurait rien de gênant s'il ne conduisait à mener une description à la fois partielle et insuffisante de la fonction et du rôle des mots en indexation. En effet, nous pourrons montrer (en II) que l'ensemble des fonctions attribuées au descripteur ne relèvent pas toutes de leur nature lexicale, qu'il faut invoquer d'autres mécanismes langagiers, d'autres niveaux, pour décrire les faits d'indexation.

L'une des marques les plus perceptibles du modèle lexicaliste en indexation se trouve dans le mode de dénomination qu'elle met en œuvre. L'indexation, qui ne pose pas la question de la signification lexicale, emprunte le modèle ancestral du baptême. Ainsi, lors de la constitution d'un langage documentaire, procède-t-on à un nouveau baptême des mots de la langue « naturelle » : on décide, par exemple[3], de « nommer », par « circulation verticale », « escalier mécanique », « ascenseur », et « monte-charge ». Ce mode de nomination, s'il est caractéristique des unités linguistiques de type « nom propre[4] », est appliqué en indexation à tous les types de nom, y compris les noms communs ; nous y reviendrons.

I.2.2 - UN MODÈLE OBJECTIVISTE DU LANGAGE

A - Approche du modèle et rappel des critiques dont il fait l'objet

Dans le modèle objectiviste, le langage est appréhendé comme un instrument au service d'une fonction : la communication. Il est en cela un mode de codage, un code.

Il ne manipule qu'un seul type de donnée : l'information, qui existe en dehors de la langue et du langage, prête à être codée. L'information étant ce qui peut être transmis, elle peut être de nature différente : concept, sens, signification, connaissance, etc., ne sont pas, dans ce modèle, distinguables.

La notion d'information ne permet pas donc pas d'établir de différences entre les mots et les textes : les textes sont appréhendés comme des ensembles de mots, des supports d'information.

---

[1] Même si, comme le relève Marandin, il n'y a aucun consensus sur ce qu'est la signification ; Marandin 1992a [présentation], p. 8 : « Alors qu'il n'y a aucun consensus sur le mode d'individuation par la signification (parce qu'il n'y a pas de consensus sur ce qu'est la signification), il semble que tout le monde s'accorde pour voir dans la signification un principe absolu d'individuation ».

[2] Ou plutôt devenu standard au fil des siècles, voir Mounin 1963, p. 25-27.

[3] L'exemple est repris de Chaumier 1978, p. 34.

[4] Voir ci-après le chapitre III, § III.2.1.

Les trois principaux présupposés du modèle objectiviste ont été largement remis en question par les théories linguistiques[1] qui posent au contraire que[2] :
(i) la langue n'est pas un instrument ;
(ii) si elle permet de communiquer, cette utilisation n'est qu'une parmi d'autres[3] ;
(iii) la langue n'est pas un code : son « contenu » ne peut à ce titre en être dissocié.

Les critiques dont le modèle objectiviste a pu faire l'objet n'empêchent pas, bien entendu, que ses présupposés restent valides pour les sujets parlants et ce d'autant plus que ce modèle de la langue comme moyen de communication bénéficie d'une longue et estimable tradition[4]. Là encore, il est tout à fait naturel que l'indexation emprunte les présupposés de ce modèle ambiant ; mais là encore, l'inscription dans ce modèle mérite d'être révisée dès lors que son incidence dépasse la simple référence culturelle implicite. En effet, les présupposés de ce modèle constituent, pour la pratique de l'indexation, d'importants facteurs d'opacification qui tendent à l'enfermer dans une impasse.

B - Les marques du modèle objectiviste en indexation

Le modèle objectiviste, adopté « inconsciemment » en indexation, se trouve exprimé dans le discours classique sous deux formes différentes :
(i) sous une forme positive : les présupposés du modèle objectiviste permettent de définir les caractéristiques du descripteur ;
(ii) sous une forme négative : les limites que l'indexation rencontre dans l'utilisation qu'elle veut faire de la langue tiennent des limites propres aux présupposés du modèle objectiviste.

(i) Les caractéristiques du descripteur décrites dans les termes du modèle objectiviste

Certaines des caractéristiques du descripteur que nous avons précédemment dégagées relèvent typiquement du modèle objectiviste, notamment les caractéristiques 1, 2 et 3 :
- le fait que le descripteur soit posé comme l'expression linguistique de concepts préexistants relève du présupposé objectiviste concernant l'extériorité du « contenu » véhiculé par la langue ;
- l'appréhension du descripteur comme une expression linguistique stable tient, elle, de l'hypothèse de la langue comme code ;
- la caractéristique du descripteur comme condensateur textuel s'inscrit, elle, dans le cadre d'une approche élargie et extensible de la notion d'information qui, dans le modèle objectiviste, annule les différences de niveaux texte/mot.

---

[1] Par exemple Chomsky [1975] 1981, mais il n'est pas, loin de là, le seul.
[2] On reprend la synthèse des critiques que propose Rastier (1994, n. 1, p. 17-18).
[3] « M. Halle a raison de s'élever contre l'attitude de ceux qui, donnant à la formule "une langue est un instrument de communication", l'interprétation extrapolée "une langue est un instrument parfait de communication" et, constatant qu'il n'en est rien, en prenant l'exact contre-pied, en une formule plus contestable encore : la langue n'est pas un moyen de communication ». Kerbrat-Orrechioni 1980, p. 12.
[4] Que la langue soit un instrument de communication et qu'à ce titre elle doive être un parfait moyen de communication se situe par exemple au cœur du projet de « langue universelle » de Descartes ; il y a, dans l'histoire, bien d'autres exemples de cette quête mythique d'une langue parfaite ; on peut sur ce point consulter Eco 1994.

Quelle conclusion tirer du constat que certaines des caractéristiques du descripteur reposent sur des représentations linguistiques non valides d'un point de vue théorique sur la langue ?

Nous sommes devant l'alternative suivante :
- soit ces caractéristiques décrivent de façon adéquate le descripteur : elles devront être maintenues, sous une autre forme, dans un autre cadre, si on cherche à les constituer en fondements ;
- soit ces caractéristiques sont essentiellement liées aux formes du modèle objectiviste : elles disparaissent alors avec lui.

Nous montrerons que, sur les cinq caractéristiques du descripteur dégagées précédemment, trois ne se justifient que dans le cadre objectiviste. On verra en effet que les caractéristiques du descripteur comme expression de concept, comme expression stable et comme condensateur textuel, caractéristiques propres à réaliser la fonction 1 de représentation du contenu d'un document, ne rencontrent aucune des propriétés de langue telles qu'un modèle linguistique peut les décrire. En revanche, ces trois caractéristiques peuvent, dans le cadre d'une théorie de la langue, recevoir un autre type d'explication, en termes d'effet d'interprétation. En ce sens, on dira que la première des fonctions de l'indexation – la représentation du contenu d'un document – relève d'un effet d'interprétation construit à partir des formes lexicales qui, rétroactivement, se voient investies du rôle de représenter ce contenu.

C'est sur ce dernier point que le modèle de fonctionnement du lexique en indexation articule les deux représentations lexicaliste et objectiviste : comme toute la langue se comprend à travers ses formes lexicales, tout ce qu'on suppose à son sujet se trouve investi au niveau des mots.

(ii) Les limites du modèle objectiviste en indexation

Après d'autres[1], on peut relever que les limites imputées à la langue dans le discours classique relèvent des limites propres aux présupposés du modèle objectiviste :

- la langue comme code : le fait que la langue soit appréhendée comme un code conduit les professionnels de l'indexation à dire qu'elle est un « mauvais » code (trop ambigu, trop irrégulier, etc.) mais qu'il est possible d'améliorer ce code moyennant quelques aménagements (réduction du nombre de formes lexicales, explicitation des relations sémantiques entre ces formes, etc., dans les langages documentaires) ;

- la langue comme véhicule d'information préexistante : ce présupposé invite les professionnels de l'information à faire preuve de grande sévérité envers ce qu'ils nomment la variabilité de l'indexation[2]. Non seulement on déplore que deux indexeurs ne puissent indexer de façon identique un même document, mais on s'étonne aussi que le même indexeur indexe différemment le même document en T_1 et en T_2. C'est là un problème que la pratique a du mal à résoudre. Comme celui de l'évolution des connaissances : la pratique de

[1] Par exemple, Turner 1990, Dubois 1995.
[2] Cf. par exemple, Bertrand-Gastaldi 1986, p. 11-13.

> l'indexation est prise dans une éternelle spirale du « retard » qui rend obsolètes, à peine achevés, les langages documentaires qu'elle met au point. Là encore, c'est à la vision d'une stabilité des connaissances et de l'information, propre au modèle objectiviste, que se heurte la pratique de l'indexation.

Certains voient, dans l'adoption implicite du modèle objectiviste du langage, des contradictions encore plus importantes, touchant la raison d'être même de l'indexation comme moyen de mise à disposition des connaissances ; ainsi Turner se montre-t-il particulièrement virulent : « L'ambiguïté est inhérente au langage. Elle est en grande partie responsable du renouvellement constant des problématiques scientifiques. Vouloir figer la signification des mots par l'adoption d'un symbolisme rigoureux est un non-sens. Un tel projet ne tient aucun compte de la dynamique de la construction sociale des connaissances.[1] »

On trouve une semblable critique chez d'autres auteurs, notamment Dubois[2].

Là encore quelle conclusion doit-on tirer de l'adoption implicite du modèle objectiviste dans l'approche du lexique en indexation ?

Il nous semble que l'on doit tenter de s'affranchir de ce modèle, non seulement parce que, n'étant pas valide d'un point de vue théorique, il ne permet pas de dégager des fondements de l'indexation, mais aussi parce que les représentations linguistiques qu'il véhicule enferment la pratique de l'indexation dans de faux problèmes[3] et l'empêchent de penser ses propre objets : document, information, descripteur, etc.

En outre, concernant précisément le descripteur, le modèle objectiviste s'avère insuffisant pour expliquer certaines de ses caractéristiques, et notamment les caractéristiques 4 et 5 :

- son autonomie : sur quoi repose, dans le modèle objectiviste, le fait que le descripteur puisse être utilisé seul, détaché du document auquel il renvoie ? Sur une autonomie conceptuelle ?

- son rôle de relais textuel : comment le descripteur, expression univoque d'un concept stable, pourrait-il renvoyer à différents concepts issus de plusieurs documents ? Si le descripteur ne doit exprimer qu'un seul et unique concept, toujours le même, comment peut-il, en tant que terme préférentiel, rendre compte aussi des concepts de ses quasi-synonymes ?

---

[1] Turner 1990, p. 2.
[2] Dubois 1995, p. 92 : « Elle [l'ambiguïté] est simplement normale, voire même « naturelle », dans la nature des langues comme de toutes les productions humaines [...] elle est en elle-même productive des connaissances. Tenter de réduire systématiquement ces écarts, plutôt que de les repérer et de les gérer comme tels, ferme en effet la possibilité d'évolution tant individuelle que collective des connaissances ».
[3] Notamment celui des langages documentaires : les langages sont toujours imparfaits et incomplets, mais, tant que l'on ne disposera pas de langages parfaits et complets, l'indexation restera d'application délicate.

I.3 - Conclusion et résultats intermédiaires

Dans notre étude du modèle d'utilisation du lexique en indexation (I.1), nous avions dégagé, à partir des deux fonctions de l'indexation, cinq caractéristiques :

- **fonction 1 de l'indexation** : « indiquer brièvement, sous forme concise, la teneur du document »
  fonction 1 du descripteur : le descripteur comme unité de représentation du contenu d'un document :
-> caractéristique 1 : le descripteur comme expression linguistique d'un concept préexistant, stable, unitaire (« simple ») ;
-> caractéristique 2 : le descripteur comme expression linguistique stable ;
-> caractéristique 3 : le descripteur comme condensateur textuel.

- **fonction 2 de l'indexation** : « permettre la recherche efficace des informations contenues dans un fonds documentaire »
  fonction 2 du descripteur : le descripteur comme accès à un fonds documentaire :
-> caractéristique 4 : le descripteur comme expression linguistique autonome ;
-> caractéristique 5 : le descripteur comme relais textuel.

Les caractéristiques 1, 2, 3 du descripteur qui relèvent de la fonction 1 de l'indexation (représentation du contenu d'un document) nous sont apparues très déterminées par l'adoption implicite du modèle objectiviste du langage. Nous avons fait l'hypothèse que ces trois caractéristiques du descripteur n'étaient propres qu'au modèle objectiviste et qu'elles n'étaient plus pertinentes dans un autre cadre. En effet, la fonction dont elle relève – celle de représenter le contenu d'un document – peut être analysée en termes d'effet d'interprétation qui, construit à partir des formes lexicales, se trouve *a posteriori* attaché aux unités mêmes, aux descripteurs.

Nous montrons ci-après, dans le point II.1, comment on peut envisager la construction d'un tel effet.

L'indexation est donc, à ce stade de notre recherche, vidée de la notion d'analyse et de représentation du contenu et se voit recentrée autour de sa fonction d'accès à l'information. Sur ce dernier point, on a pu remarquer que les caractéristiques 4 et 5 du descripteur, qui relèvent de la fonction 2 de l'indexation, ne peuvent être traitées de façon adéquate dans le cadre du modèle objectiviste.

Dans le modèle de fonctionnement du lexique que nous proposons ci-après, et dans lequel nous mènerons notre étude de la représentation du contenu comme effet, les caractéristiques 4 et 5 du descripteur pourront être analysées et proposées comme fondements théoriques.

II - Déplacement du modèle de fonctionnement du lexique

Dans ce paragraphe, nous étudierons les cinq caractéristiques du descripteur précédemment dégagées au regard de deux modèles de description linguistique des

unités lexicales. Notre objectif est de distinguer, parmi les caractéristiques du descripteur, celles qui relèvent d'effets propres à une utilisation de la langue et celles qui relèvent de principes propres à la langue.

Compte tenu des deux plans mis en jeu dans l'indexation (plan des textes et plan des mots), nous travaillerons à partir des deux modèles linguistiques suivants :

- pour comprendre le rapport que l'indexation établit entre texte et mot, nous étudierons ce que le *modèle de l'analyse de discours* dit du rapport entre thème et discours ; c'est sous cet angle que nous essayerons de distinguer faits et effets dans les caractéristiques 1, 2, 3 et 5 du descripteur (II.1) ;

- pour approcher le « saut » qu'opère l'indexation, par le biais du mot, entre « contenu » d'un document et « information » d'une collection documentaire, nous examinerons le statut des unités lexicales dans le cadre d'*une théorie du lexique** : cette approche linguistique des unités lexicales hors emploi nous permettra de préciser les caractéristiques 4 et 5[1] du descripteur (II.2).

Le déplacement de modèle de fonctionnement que l'on propose ici se marque par l'introduction, dans ce paragraphe, de nouvelles notions, proprement linguistiques[2], pour lesquelles le glossaire situé en fin d'étude propose une première approche.

II.1 - Distinction des faits et des effets par le biais du modèle de l'analyse de discours

Sous une forme très simplifiée, nous exposons dans ce paragraphe quelques éléments issus du modèle linguistique de la « nouvelle analyse de discours[3] ». Ce modèle, originellement proposé par Pêcheux dans les années 1970 en réponse aux insuffisances des approches alors dominantes[4], a été particulièrement travaillé par Marandin.

L'approche de Marandin permet de montrer comment se construisent entre mots et textes des effets d'interprétation : des effets de « cohérence thématique » quand un mot est identifié comme thème dans un discours et des effets d'« intertextualité[5] » quand le mot d'un discours semble appeler ou rappeler d'autres discours. Marandin montre en effet que le thème de discours se construit nécessairement au travers d'autres discours, la notion de discours étant, dans ce cadre, un « construit » : « elle ne renvoie pas à la simple donnée d'un enchaînement d'énoncés[6] ».

---

[1] La caractéristique 5 du descripteur (le descripteur comme terme textuel) est abordée sous deux angles : sous l'angle du texte (§ II.1) et sous l'angle du terme (§ II.2).
[2] On verra apparaître les notions de : unité lexicale, signification lexicale, interprétation d'un texte, thème de discours, objet de discours, référent.
[3] Abrégée dans ce document sous la forme « analyse de discours ». La nouveauté de l'approche est exprimée notamment dans Marandin 1979 et 1993 et évaluée dans Marandin 1997.
[4] Notamment celles proposées par les grammaires de textes (qui analysent le texte sur le modèle d'une analyse de la phrase), dont on peut trouver une critique dans Marandin 1979.
[5] Marandin [1993] rejette explicitement le terme en raison de l'utilisation qui en est faite dans un autre cadre que le sien (cadre structuraliste) ; il propose la notion d'interdiscours, qui n'est cependant pas directement concurrente, voir ci-après, § II.1.2. Nous ne pouvons entrer dans le détail de ces discussions.
[6] Marandin 1997, p. 12.

C'est pour ces raisons que le cadre de l'analyse de discours nous paraît pertinent :

- il dégage une voie pour penser le lien entre les deux fonctions de l'indexation distinguées dans la norme : l'analyse d'un document suppose-t-elle toujours la prise en compte d'autres documents ? Dans ce cadre, comment se pense le lien entre l'analyse d'un document et l'analyse de plusieurs documents ? ;

- il fournit des pistes pour redéfinir l'indexation : comment se traduit le changement de problématique, de l'« analyse de contenu » à la constitution de l'intertextualité ?

II.1.1 - LA NOTION DE REPRÉSENTATION DU CONTENU D'UN DOCUMENT : UN EFFET D'INTERPRÉTATION

En disant que la représentation du contenu relève, en indexation, d'un effet d'interprétation, nous ne voulons pas dire que l'indexation ne permet pas, ou permet faussement, de se faire une idée du contenu d'un document. Nous tenons que l'indexation d'un document (en l'occurrence les descripteurs attribués à un document) permet bel et bien de se faire une idée du document mais nous disons que « cette idée » n'est pas liée à l'« idée », au « concept » contenu dans le descripteur, dans le mot lui-même. C'est par le biais de son interprétation en discours que le descripteur permet de rendre compte du contenu d'un document.

Pour expliquer comment des mots (les descripteurs) peuvent donner l'impression, à juste titre sans aucun doute, qu'ils disent le contenu d'un texte, nous présentons une recherche menée par Marandin[1] dans le cadre de l'analyse de discours (A). Nous aborderons ensuite (B) la question de savoir si, en indexation, un descripteur peut être pour un document ce qu'un thème est, en analyse de discours, pour un discours.

A - Le thème de discours dans le cadre de l'analyse de discours

Marandin [1988] s'intéresse à la notion de thème de discours sous l'angle de la problématique suivante :
- qu'est-ce qu'un thème de discours ?
- comment se matérialise-t-il dans le discours ?
- comment intervient-il dans la compréhension d'un texte ?

Selon lui, le point de vue sur le thème est nécessairement celui de l'interprétation (point de vue de la réception) ; la question centrale peut se résumer alors de la façon suivante : comment détermine-t-on un thème de discours ?

Ce qui nous intéresse dans la démarche de Marandin, c'est qu'elle rend compte de la formulation lexicale du thème dans des termes autres que ceux de la réduction, de la condensation d'un texte par des mots. Parallèlement, c'est une approche qui dit qu'un mot peut rendre compte d'un texte, et même que l'on ne peut rendre compte d'un texte que par des mots, qu'en le nommant et le renommant[2]. C'est

---

[1] Marandin 1988, p. 67-87.

[2] Marandin cite à ce sujet Barthes : « Quiconque lit un texte rassemble certaines informations sous quelques noms génériques et c'est ce nom qui fait la séquence ; la séquence n'existe qu'au moment où et parce qu'on peut la nommer, elle se développe au rythme de la nomination qui se cherche et se confirme », Barthes 1970, p. 14 (cité *in* Marandin 1997, p.

ainsi que cette approche peut permettre de comprendre comment texte et mot peuvent être assimilés, confondus, et comment l'on peut *a posteriori* désigner le mot comme porteur de l'interprétation d'un texte.

Avant d'en venir à l'explication du mécanisme de l'assimilation thème/discours, nous présentons ci-après les principales propositions de Marandin :

(i) un thème, c'est toujours un *nom*, et plus précisément, un individu linguistique de type syntagme nominal (abrégé SN ci-après, appelé aussi GN, groupe nominal, par Marandin) ;

(ii) un thème, c'est un nom qui a la *capacité d'organiser* un texte : cette approche du thème s'oppose à celle qui voit dans le thème un nom qui dit ce « à propos de quoi » est le texte[1] ;

(iii) organiser un texte, c'est en donner une interprétation[2] : le thème est à ce titre un SN apte à *projeter une interprétation* ; on remarque là encore que, dans cette approche, le thème n'est pas un nom qui interprète un texte ou une partie de texte.

La forme lexicale d'un thème de discours n'est donc pas une interprétation de ce discours mais plutôt un élément déclencheur de la construction d'une interprétation.

Thématiser revient alors à constituer des unités d'interprétation et non à retrouver dans un texte des unités déjà interprétées (les éléments d'une encyclopédie dans les termes de Marandin[3]).

C'est à ce titre que le thème est nécessairement tissé à même la trame du texte : c'est toujours un terme textuel, un terme par rapport à un texte, que le nom du thème soit issu du texte ou qu'il soit déterminé de l'extérieur, dans l'espace de réception du texte[4]. Dans les deux cas, c'est le texte qui dispose de quoi constituer le thème.

C'est cette caractéristique du thème comme terme textuel qui est à l'origine de l'« illusion d'optique », de l'effet d'eschérisation dit encore Marandin, que génère le thème : « La possibilité pour un groupe nominal de désigner un individu dans un

---

21). Dans Marandin 1988, l'article s'ouvre sur cette autre citation, proche de la précédente dans l'esprit, de Barthes : « Lire, c'est trouver des sens, c'est les nommer ; mais ces sens nommés sont emportés vers d'autres noms ; les noms s'appellent, se rassemblent et leur groupement veut de nouveau se faire nommer : je nomme, je renomme ; ainsi passe le texte : c'est une nomination en devenir, une approximation inlassable, un travail métonymique », Barthes 1970, p. 17 (cité *in* Marandin 1988, p. 67).

[1] Voir Marandin 1997, p. 21 : « Le thème de discours est un aspect du processus de compréhension et non pas la donnée d'un individu externe à propos de quoi le discours se tient ».

[2] Un thème est en cela une manière de comprendre. Il y a toujours plusieurs manières de comprendre et donc plusieurs thèmes possibles ; la question reste de « déterminer comment une même suite d'énoncés permet des lectures différentes et ce qui, dans les énoncés ou leurs enchaînements, oriente vers telle lecture plutôt que vers telle autre ». Marandin propose des pistes, 1988, p. 71 et suiv.

[3] *Ibid.*, p. 84.

[4] C'est la différence entre thème configuré et thème inféré qu'établit Marandin 1988 p. 77 et suiv. Nous l'évoquons ci-après.

monde, quand il est traité ou interprété dans un énoncé d'occurrence, et le monde dans lequel s'identifie cet individu, quand il est traité comme thème de discours. »[1]

C'est ce mécanisme d'identification entre individu et monde, c'est-à-dire entre thème et discours, que nous allons essayer d'expliquer en reprenant l'analyse et les exemples de Marandin.

Le point de l'analyse porte sur l'approche du terme textuel comme un terme par rapport à un texte. Ce qui fait un terme textuel, c'est le « contenu descriptif » d'un terme. Un contenu descriptif est relatif à un discours donné : il correspond à ce qui est introduit par les énoncés et ensuite transformé en propriétés du terme.

Soit l'exemple suivant repris de Marandin 1988 : « La licorne à fourrure d'hermine abondait autour du château. Un jour, Lancelot s'amusa à les pourchasser. Piqué au jeu, il les tua toutes. Puis, il les dépouilla et il s'empara de leur précieuse toison. Trois jours après, il mourrait dans d'affreuses douleurs. Lancelot fut pleuré de tous. Isolde s'enferma dans un couvent... »

Dans cet exemple, on peut dégager deux types de thème : un thème configuré (nommé dans le texte) et un thème inféré (nommé dans l'espace de réception du texte).

Comme thème configuré, il y a par exemple *Lancelot*, dont le contenu descriptif est /celui qui s'appelle Lancelot, qui a pourchassé, tué, dépouillé les licornes et qui est mort après/. Si *Lancelot* est le thème de ce discours, on dira que ce discours n'est pas « à propos de » Lancelot mais « à propos de » /celui qui s'appelle Lancelot, qui a pourchassé, tué, dépouillé les licornes et qui est mort après/, etc.

Comme thème inféré, il y a par exemple *la vengeance des licornes*, dont le contenu descriptif, « abstrait de la suite d'énoncés[2] », peut être /la mort de Lancelot est liée à ce qu'il a fait trois jours auparavant, les licornes ont causé la mort de Lancelot, les licornes font mourir Lancelot pour venger un méfait dont elles sont les victimes, leur extermination étant un méfait à leur encontre/. Là encore, si *la vengeance des licornes* est un thème du discours, on dira que ce discours n'est pas « à propos de » la vengeance des licornes mais « à propos de » /la mort de Lancelot est liée à ce qu'il a fait trois jours auparavant, les licornes ont causé la mort de Lancelot, les licornes font mourir Lancelot pour venger un méfait dont elles sont les victimes, leur extermination étant un méfait à leur encontre/.

Le contenu descriptif du terme qui en fait un terme textuel et par suite un thème de discours peut donc être défini comme « un agrégat subsumant d'autres individus dans leurs interrelations, telles qu'elles sont *introduites dans les énoncés, reconstruites dans la compréhension et constitutives d'une interprétation*[3] ».

Cette approche du thème de discours comme terme textuel nous paraît précieuse en ce qu'elle fait apparaître deux dimensions de la thématisation :

---

[1] Marandin 1988, p. 82.
[2] Marandin 1997 [p. 23] met l'accent sur ce qui permet dans un texte d'inférer un thème : ce sont les transitions temporelles ; pour une description, voir Marandin 1988, p. 81.
[3] Marandin 1988, p. 82.

- une dimension de l'interprétation[1], qui affaiblit l'idée qu'un texte aurait, hors lecture, en soi, un thème, des thèmes et que, hors lecture, en eux-mêmes, ces thèmes pourraient dire un contenu ;

- une dimension de la transformation : la transformation d'énoncés en propriétés permet de comprendre comment un terme isolé de son contexte peut donner l'impression de dire le contenu d'un texte, ou d'avoir un contenu tout court.

Cette analyse du thème de discours qui démonte le mécanisme de l'assimilation thème/discours peut-elle être reprise dans le cadre de l'indexation ?

B - Le thème de discours en indexation : première approche[2]

Pour que la recherche de Marandin prenne un sens dans notre cadre, il faut que le descripteur puisse être considéré comme un thème de discours.

C'est là une proposition qui n'est pas nouvelle en indexation. En effet, Michel Le Guern[3] a pu défendre que :
- le descripteur est une unité de discours ;
- le descripteur est, plus précisément, un syntagme nominal ;
- l'indexation peut être vue comme la détermination de thèmes[4].

L'hypothèse du descripteur comme thème de discours est donc déjà posée ; nous la reprenons à notre compte dans cette recherche et nous donnons pour objectif de la préciser, notamment en examinant :
- comment un document peut être considéré comme un discours, objet du chapitre IV ;
- comment (et par qui) le descripteur se constitue au fil du discours, à même le texte, comme un terme textuel, objet du chapitre V.

Reste qu'adopter l'hypothèse du descripteur comme thème de discours n'est pas, même dans le cadre d'un modèle d'utilisation (d'un modèle qui utilise le concept linguistique de thème), sans poser des problèmes, notamment de représentativité du phénomène : nous en abordons deux ci-après.

(i) Tous les descripteurs peuvent-ils être perçus comme des thèmes de discours ?

À simplement regarder une formule d'indexation telle que la produisent les organismes documentaires, on remarque que tous les descripteurs utilisés ne peuvent fonctionner comme thèmes de discours. Ainsi dans cette indexation[5]

---

[1] Nous n'avons pas vraiment développé dans ce paragraphe cet aspect de la recherche de Marandin [1988, p. 70], c'est-à-dire son idée que « la problématique du thème de discours recoupe celle de l'anaphore comprise au sens de "dépendance interprétative" que contractent entre eux les GN d'un texte ». Elle repose sur la notion de chaîne de référence, que nous étudierons au chapitre V.
[2] Une approche plus approfondie est proposée dans le chapitre V consacré au descripteur.
[3] Le Guern 1984, notamment.
[4] *Ibid.*, p. 168.
[5] Voir annexe 1 : les données de notre enquête.

réalisée par la *Documentation française* à partir d'un article du *Monde*[1] : « Tapie Bernard, Bernard Tapie Finances, Groupe Bernard Tapie, tribunal de commerce, jugement, redressement judiciaire, délai, personnalité position, Crédit Lyonnais, affaire, navire, fortune, cour d'appel, Testut, amende, sanction, COB. »

On remarque que – et c'est là une particularité des langages documentaires qui tend à opacifier le mécanisme de l'indexation quand elle y a recours – tous les descripteurs utilisés en indexation ne jouent pas le même rôle par rapport au document : tous ne décrivent pas le « contenu », tous ne sont pas des « descripteurs matières ». C'est le cas de ceux que l'on appelle les « mots-outils[2] » (dans l'exemple : « affaire » et « délai ») ou encore les « mots-facettes[3] » (dans l'exemple : « personnalité position »).

L'indexation effectuée avec le langage documentaire Rameau[4] est sur ce point tout à fait exemplaire. Une « vedette-matière » (la formule d'indexation complète) est constituée systématiquement d'éléments de nature hétérogène :

- une « tête de vedette » qui correspond à un mot (ou plusieurs) qui « exprime » le « sujet » d'un document : par exemple « cerveau » ;

- des subdivisions, qui sont des mots chargés de préciser la tête de vedette sous différents points de vue : point de vue du sujet (par exemple « maladies » pour le sujet « cerveau ») ; point de vue géographique (par exemple « Italie ») ; point de vue chronologique (par exemple « Renaissance ») ; point de vue de la forme du document (par exemple « répertoires »).

Ainsi, dans la formule d'indexation suivante issue de Rameau – « cerveau**maladies**Italie**Renaissance**répertoires[5] » –, chacun des mots est à lire et à interpréter d'une façon particulière, qui, certes, est spécifiée dans les gros volumes[6] qui accompagnent le répertoire du vocabulaire Rameau, mais qui ne présente aucun caractère d'évidence.

Il y a sans aucun doute, sur le point précis des langages documentaires, toute une série d'analyses de détail à mener qui permettraient de mettre au jour d'autres aspects de la perception du lexique chez les professionnels de la documentation. Pour ce qui concerne notre sujet de recherche – l'indexation –, il nous paraît essentiel de pouvoir prendre des distances par rapport aux présupposés propres à tel ou tel langage documentaire : nous considérons la notion de langage documentaire dans son principe et, à ce titre, nous constatons qu'elle constitue une difficulté pour concevoir le descripteur de façon unifiée (par la notion de thème de discours). Mais cette difficulté ne constitue pas pour autant un obstacle, notamment si on peut montrer, ne serait-ce que partiellement, que l'indexation peut réaliser son objectif en se passant de l'usage d'un langage documentaire. L'importation de

---

[1] *Le Monde* du 1/12/1994, p. 24.
[2] « Descripteur n'ayant pas de valeur documentaire spécifique et généralement utilisé en association avec un ou plusieurs descripteurs », norme Z 47-100 (1981), p. 19.
[3] « Catégories de notions de même nature tels que processus, phénomène, matériau, outil, permettant un regroupement de termes indépendamment des disciplines traitées ». *Id.*
[4] Répertoire d'autorité matière encyclopédique et alphabétique.
[5] Pour des raisons peu claires, le langage Rameau préconise l'usage du pluriel, « sauf quand l'usage l'interdit » ; les exemples donnés sont tous des noms propres : « jugement dernier », « communion des saints », « France ».
[6] Guide d'indexation RAMEAU [Bibliothèque nationale de France 1995].

la notion de thème de discours dans un modèle d'utilisation du lexique en indexation n'est donc pas nécessairement entravée par les modes de fonctionnement particuliers que mettent en œuvre les langages documentaires. Si elle est problématique, c'est essentiellement dans un cadre où l'indexation ne se pense qu'au travers d'un langage documentaire ; or cette position empêche de définir en propre le descripteur qui – on l'a vu – se donne sous une forme singulièrement hétérogène.

(ii) Tous les aspects du descripteur peuvent-ils être rendus par la notion de thème de discours ?

Un autre type de problème peut faire apparaître l'importation de la notion de thème de discours en indexation comme sévèrement restrictive. En effet, l'indexation doit, comme nous l'avons vu, permettre d'accéder à l'information d'un fonds documentaire et, à ce titre, le descripteur doit pouvoir être appréhendé à la fois comme une unité autonome et comme un synonyme (fonction 2 de l'indexation) : nous aurons donc à réinterroger la notion de thème de discours sous l'angle de ces deux aspects du descripteur[1].

Ce paragraphe (II.1.1) avait pour objet de préciser comment s'établit la notion de représentation de contenu en indexation.

L'analyse du thème de discours proposée par Marandin montre comment un thème peut représenter un discours : par le biais d'une illusion d'optique déclenchée sur la base du contenu descriptif d'un terme compris comme terme textuel. Si l'on adopte l'hypothèse que le descripteur peut se comprendre, dans le cadre d'un modèle d'utilisation du lexique, comme un thème de discours, alors on peut poser que l'analyse de contenu en indexation relève de cette même illusion d'optique, d'un effet déclenché sur la base des descripteurs par le biais d'une interprétation : nous y reviendrons.

Cette approche nous conduit à voir que la notion de représentation de contenu en indexation, si elle relève de l'interprétation (donc du niveau de la réception), ne peut être donnée pour un principe de production : si les descripteurs en tant que termes textuels donnent la possibilité de représenter le contenu, ils ne sont pas eux-mêmes des unités de représentation de contenu.

Cherchant à montrer que la représentation du contenu relevait en indexation d'un effet d'interprétation, nous avons du même coup posé une nouvelle hypothèse et ouvert une nouvelle piste de recherche :

- *l'hypothèse du descripteur comme thème de discours, hypothèse qui nous éloigne de la conception objectiviste d'une transmission stabilisée de concept, de l'auteur à l'utilisateur : il apparaît en effet que le descripteur intéresse moins pour son « contenu » propre (conceptuel ? sémantique ?) que pour le contenu textuel (le « contenu descriptif ») auquel il permet d'accéder ;*

- *cette hypothèse engage la recherche dans des voies nouvelles : qu'est-ce que l'indexation met à disposition de l'utilisateur pour qu'il construise son interprétation ? En quoi les accès aux documents qu'elle dispose sont-ils à même de permettre la construction de ces interprétations ? L'approche du*

[1] Voir, sur ce point, le chapitre V.

descripteur comme thème de discours est-elle suffisante pour caractériser l'ensemble du fonctionnement attendu du descripteur ?

C'est à travers les données de l'analyse de discours que nous avons proposé de comprendre la notion de représentation de contenu en indexation comme un effet, un effet d'interprétation qui se donne, dans le discours classique, pour un principe de production.

C'est à nouveau par l'analyse de discours que nous examinerons comment le descripteur peut fonctionner comme un relais textuel, c'est-à-dire comment un mot peut rapprocher plusieurs textes différents.

II.1.2 - LA NOTION D'ACCÈS À UN FONDS DOCUMENTAIRE : UN PRINCIPE D'INTERPRÉTATION EN ANALYSE DE DISCOURS

Nous avons précédemment dégagé du modèle d'utilisation du lexique en indexation cinq caractéristiques du descripteur, réparties en deux groupes correspondant aux deux principales fonctions de l'indexation. Parmi ces caractéristiques, se trouve celle du descripteur comme relais textuel, c'est-à-dire comme forme lexicale susceptible de lier entre eux plusieurs documents différents. Nous avions noté que le modèle objectiviste du langage se révélait insuffisant pour décrire à la fois la possibilité et le fonctionnement de cet aspect du descripteur. Le cadre d'analyse que nous avons précédemment adopté, celui de l'analyse de discours, nous permet de traiter cette caractéristique du descripteur, et donc de traiter l'ensemble les deux fonctions de l'indexation dans les mêmes termes : c'est sur cette base que nous pourrons avancer une hypothèse qui explique le lien entre les deux aspects de l'indexation[1].

Si le paragraphe précédent traitait de l'écart qui maintient distincts thème/discours et des effets interprétatifs qui peuvent les lier, celui-ci traite de l'écart qui maintient distincts $discours_1$/$discours_2$...$discours_n$ et des effets interprétatifs qui, toujours à partir d'un mot, peuvent les mettre en relation. Si les problématiques ne sont pas exactement les mêmes, elles présentent des points communs. Les deux reposent sur l'appréhension des formes lexicales comme « relais », mais à deux niveaux différents :

- la relation mot/texte engage la problématique à un niveau de l'intradiscours ;
- la relation texte/texte par le mot l'engage à un niveau de l'interdiscours.

A - Les notions d'interdiscours et intradiscours en analyse du discours

L'un des traits majeurs du courant de l'analyse de discours dans lequel s'inscrit Marandin consiste à établir une relation fondamentale entre deux niveaux : l'intradiscours, l'analyse d'un discours, ne tient que par l'interdiscours, l'analyse à travers plusieurs discours. L'enjeu peut être ainsi formulé : « On peut bien dire que *l'intradiscours en tant que le "fil du discours" du sujet est strictement un effet de l'interdiscours sur lui-même*, une "intériorité" entièrement déterminée comme telle de l'"extérieur".[2] »

---

[1] On se souvient que, dans la norme, le lien entre les deux fonctions n'est pas spécifié (*supra*).

[2] Pêcheux 1975, cité *in* Marandin 1997, p. 12. C'est nous qui soulignons.

La présence d'autres discours dans un discours se marque par les mots, qui n'entrent jamais « seuls » dans les énoncés, mais qui ouvrent au contraire « la mémoire et l'anticipation d'autres textes[1] ». Le thème discursif, se construit au fil du texte mais dans un texte, il y a toujours des « hiatus sémantiques », des espaces où peuvent se glisser des bribes d'autres textes.

Dans ce type d'analyse de discours, un mot n'est jamais, à proprement parler, mis face à un texte. L'analyse du thème de discours ne relève en effet pas d'une sémantique lexicale, mais d'une sémantique discursive[2] : « On retrouve ici le point de départ historique de l'analyse de discours : une réflexion sur le mot dans la lexicographie politique, puis dans la critique de la sémantique lexicale effectuée par Pêcheux dans *Les Vérités de la Palice*. Mais la problématique s'est déplacée d'une approche sémiotique où le mot est élément de système (de langue ou idéologique) à une problématique sémantique où il est marqué (chargé dirait Bakhtine) par ses occurrences dans d'autres textes (les usages qui en sont faits ou les textes qui s'y condensent).[3] »

Si toute analyse de discours suppose l'analyse de plusieurs discours, les mots du discours analysé sont nécessairement analysés aussi dans d'autres discours. Dès lors qu'un même mot, qu'un même thème de discours, puisse renvoyer à des discours différents ne surprend plus puisqu'il n'a été constitué comme tel qu'à partir d'une approche interdiscursive.

Si donc le descripteur peut fonctionner, dans un cadre qui reste à définir, comme thème de discours, alors il fonctionne nécessairement comme relais textuel : ce n'est là qu'un autre effet, quoiqu'à un autre niveau, du fait qu'il soit thème de discours.

B - Les notions d'interdiscours et d'intradiscours en indexation : première approche[4]

On voit donc que les caractéristiques du descripteur, si on le considère comme thème de discours, sont moins divergentes qu'il n'y paraît. Que le descripteur puisse être considéré comme représentant le contenu d'un texte (fonction 1 de l'indexation) comme celui de plusieurs textes qu'il met en relation (fonction 2) relève pareillement d'effets interprétatifs, de nature différente, mais réalisés sur la même base : celle de la construction du thème de discours, compris dans un cadre où l'intradiscours suppose l'interdiscours. Du coup, la question qui concerne en propre l'indexation devient celle de la production, de la construction de ces effets d'interprétation : comment l'indexation permet-elle de créer des thèmes de discours qui se définissent par leur cheminement à travers plusieurs discours ?

On voit que, si le descripteur est un thème de discours, l'ensemble des discours (des textes, des documents) qui permettent de le constituer comme tel prend une importance déterminante : à ce titre, l'indexation pourrait être comprise comme une construction de l'interdiscours[5].

---

[1] Marandin 1984, p. 53.
[2] C'est exactement la position que défend Le Guern [1984, 1991a par exemple] pour l'approche du descripteur en indexation : nous y revenons en III.
[3] Marandin 1984, p. 53.
[4] Une approche plus approfondie est proposée dans le chapitre IV.
[5] C'est une hypothèse que nous défendons dans le chapitre IV.

En effet, il est clair qu'un thème de discours ne se construit pas à travers n'importe quels discours ; il est nécessaire de constituer un ensemble de discours susceptibles de s'entrecroiser : « On peut se donner pour objectif de suivre la trace [des] textes présents, absents dans le texte de départ. Il faut alors définir une *stratégie d'exploration*. Diverses stratégies sont concevables. L'une d'elle se définit sur le modèle co-textuel (...) : des fragments de textes sont rassemblés parce qu'ils se donnent dans les formes que l'analyse intra-textuelle décrit comme semblables.[1] »

Si le descripteur est un thème de discours, un thème issu de l'interdiscours, l'indexation porte donc non plus sur un document mais sur plusieurs documents, un fonds documentaire ; et l'essentiel de sa stratégie consiste moins à déterminer le contenu d'un document qu'à déterminer des principes de regroupement des documents. À la suite de Marandin en analyse de discours, nous proposerons des principes de regroupement qui empruntent leurs formes aux « formations discursives » proposées par Foucault (*infra* chapitre IV).

II.1.3 - CONCLUSION ET RÉSULTATS INTERMÉDIAIRES

Les questions qui se posaient au début du paragraphe II de ce chapitre étaient les suivantes : les caractéristiques du descripteur dégagées de notre lecture de la norme sont-elles définitoires du descripteur ? Peut-on à partir d'elles constituer des fondements théoriques de l'indexation ? Que deviennent, dans le cadre d'une approche linguistique du lexique, les caractéristiques du descripteur qui semblaient si fortement déterminées par le modèle objectiviste sous-jacent au discours classique ?

Le paragraphe II.1 nous a permis de répondre à la dernière de ces trois questions et partiellement aux deux premières. Nous avons proposé de considérer les caractéristiques 1, 2 et 3 du descripteur, et plus globalement la fonction 1 de l'indexation, comme relevant d'un principe d'interprétation, considéré *a posteriori* comme un principe de production dans le discours classique : ce n'est qu'*a posteriori* et sur la base des présupposés lexicaliste et objectiviste que l'on peut attribuer aux seuls mots l'essentiel du processus de l'indexation. De la même façon, la caractéristique 5 du descripteur apparaît, à la lumière des propositions de l'analyse de discours, comme un effet d'interprétation. À ce titre, ces caractéristiques ne constituent pas des propriétés définitoires du descripteur et ne peuvent permettre d'établir les principes de l'indexation.

Dans le paragraphe suivant, nous complétons notre réponse à nos deux premières questions en étudiant la fonction 2 de l'indexation (fournir des accès au document, à plusieurs documents).

II.2 - Propriétés remarquables des unités lexicales

Notre lecture de la norme nous a permis de dégager deux caractéristiques du descripteur pour que l'indexation puisse réaliser sa fonction de fourniture d'accès à un document comme à plusieurs documents :

- Le descripteur doit révéler une certaine autonomie pour fonctionner seul, être détaché de son contexte : quel type d'autonomie est ici en jeu ? Cet aspect reste implicite dans la norme : il est donné comme une évidence. Le modèle

[1] Marandin 1984, p. 54. C'est nous qui soulignons.

objectiviste suggère que l'autonomie du descripteur pourrait être de nature conceptuelle. Or nous avons dû mettre de côté la notion de « concept », qui nous est apparue trop floue car très large : pouvons-nous penser l'autonomie du descripteur dans d'autres termes ? Pourquoi cette autonomie peut-elle être perçue de façon si évidente ?

- Le descripteur doit permettre de créer des classes d'équivalence. La norme ne dit pas clairement de quel type d'équivalence il s'agit : la notion d'équivalence lexicale semble être mise de côté, mais la notion d'équivalence documentaire, entre des documents, n'est pas explicite. Sur ce point, le modèle objectiviste, sous-jacent au discours de la norme, ne nous aide guère. Il pointe au contraire les limites d'une approche conceptuelle du descripteur : comment une expression censée représenter un seul et unique concept de façon stable pourrait-elle aussi représenter d'autres concepts, eux-mêmes uniques, de façon tout aussi stable ? En mettant de côté, là encore, la notion de concept, peut-on dégager des modes de fonctionnement du descripteur qui expliquent comment puisse se réaliser cette multi-désignation ?

C'est à partir de ces deux questions que nous examinerons la théorie des unités lexicales hors emploi : deux propriétés remarquables se dégagent, l'une concerne l'autonomie lexicale, l'autre la synonymie.

II.2.1 - LA POSSIBILITÉ D'UTILISER DES UNITÉS LEXICALES HORS EMPLOI : LA QUESTION DE L'AUTONOMIE LEXICALE

A - Approches de l'autonomie lexicale

La question de l'autonomie des unités lexicales se pose dans le cadre des théories du lexique entendues comme études des unités lexicales hors emploi[1].

Si la question de l'autonomie lexicale est pertinente dans ce contexte, elle reçoit des modes de description différents selon les approches. On peut distinguer principalement trois positions[2] :
(i) l'autonomie d'une unité lexicale est de nature référentielle, cette position est représentée par Kayser 1987 ;
(ii) l'autonomie d'une unité lexicale est de nature « opératoire », cette position est tenue par Franckel 1992 ;
(iii) l'autonomie d'une unité lexicale est de nature sémantique, cette position est défendue par Marandin 1992.

Ces trois positions sont en réalité plus intriquées qu'il n'y paraît, les deux premières se constituant sur une critique de la troisième.

---

[1] Dans le vaste champ de la recherche linguistique, la question de l'autonomie d'une unité lexicale constitue elle-même une problématique. Nous n'aborderons pas ici les arguments qui vont contre la possibilité d'étudier les unités lexicales hors emploi.

[2] Pour l'ensemble de ce paragraphe, on s'inspire du compte rendu de deux colloques organisés par le CELEX (Centre d'études sur le lexique / CNRS) ; le premier, tenu en 1990, portait sur la « définition », le second en 1992 sur l'« individualité lexicale ». On s'inspire plus particulièrement de la seconde série de travaux où trois auteurs (Milner, Franckel, Marandin) confrontent leurs positions à partir d'un même projet : la description hors emploi des unités lexicales.

Les positions (i) et (ii), qui se différencient par le mode de résolution qu'elles proposent, abordent en effet la question de l'autonomie de l'unité lexicale sous le même angle : comment rendre compte de la multiplicité des objets (*i.e.* référents) auxquels renvoie une même unité lexicale (i) ? Comment rendre compte de la multiplicité d'emplois d'une même unité lexicale (ii) ? Ces deux positions se fondent sur une critique de la signification lexicale qui, pour être communément donnée comme unique et complète, se révèle pourtant inapte à rendre compte de la multiplicité, des référents ou des emplois :

(i) pour Kayser, la « sémantique n'a pas de sens[1] » : le « sens unique» que l'on attribue généralement à une unité lexicale (la définition lexicographique par exemple) conduit à une « impasse » en ce qu'il est incapable de rendre compte de la multiplicité des catégories référentielles auxquelles peut renvoyer une unité lexicale ;

(ii) pour Franckel, un terme[2] n'a pas de signification stable : tout terme « est *a priori* susceptible de contribuer à l'émergence d'une multiplicité de valeurs sémantiques qui ne s'engendrent que par interaction avec l'environnement contextuel[3] ».

C'est ainsi que ces deux approches proposent des modes de description des unités lexicales qui se passent de la notion de sens :

(i) Kayser propose un modèle qui se fonde sur les interprétations référentielles possibles d'une unité lexicale. Par exemple, pour « livre », on dispose des choix suivants : « lecture de », « écriture de », « reliure de », etc. Les choix peuvent être effectués automatiquement, par un générateur qui explore successivement les nœuds d'un réseau ; à chaque nœud, une interprétation référentielle, dont on vérifie la compatibilité avec le contexte d'occurrence du terme analysé. Dans ce modèle, une interprétation (un nœud) est déclarée satisfaisante si elle ne débouche pas sur une contradiction ;

(ii) Franckel propose un modèle qui se fonde sur des « schèmes opératoires[4] » : « chaque terme correspond à un schème particulier, c'est-à-dire à une configuration spécifique de paramètres[5] ». Les paramètres sont principalement de deux ordres : S (sujet énonciateur) et T (espace, temps). Au terme d'une analyse dans laquelle nous n'entrons pas ici[6], on pourra déterminer le schème opératoire du terme *porter* par exemple comme étant : « la construction de la localisation de X par Y dans le temps relativement à la non-localisation de X par Y hors du plan temporel[7] ».

---

[1] Kayser 1987.
[2] Dans ce texte, Franckel [1992, p. 18] utilise le mot « terme » pour désigner « de façon délibérément vague toute entité lexico-morphologique ».
[3] *Id.*
[4] Ce type d'analyse s'inscrit dans le cadre de la théorie de Culioli, qui cherche à « concilier la notion d'invariance que suppose la notion même de propriétés intrinsèques (hors emploi) d'un terme, et celle de déformabilité et d'instabilité sémantiques à laquelle renvoie la multiplicité des valeurs que, par cette interaction, ce terme contribue en règle générale à engendrer ». *Ibid.*, p. 21-22.
[5] *Ibid.*, p. 22.
[6] On renvoie à Franckel 1992, p. 26-36.
[7] *Ibid.*, p. 34.

Il ne nous appartient pas de discuter ces modèles de description non sémantiques des unités lexicales hors emploi. En revanche, on peut relever, avec Marandin[1], que ces deux modèles posent la question de la signification lexicale dans les termes d'une alternative que l'on peut tenir pour réductrice ; en effet,

- soit les mots changent de signification selon les discours où ils sont employés : on en déduit qu'il n'y a pas de signification lexicale stable, et donc pas de signification lexicale tout court ;

- soit les mots ne changent jamais de signification : mais cette approche ne tient pas au regard des faits.

On peut aussi poser la question de l'autonomisation des unités lexicales par la signification dans un autre cadre. La signification lexicale reste un facteur individuant ; si elle ne peut rendre compte de phénomènes complexes (multiplicité référentielle ou sémantique), c'est plutôt parce qu'elle est elle-même complexe, « trop massive » : « elle recouvre des phénomènes qu'il importe de distinguer[2] ». Autrement dit, si la signification lexicale est considérée comme complexe et hétérogène[3], elle comprend alors plusieurs dimensions qui peuvent varier différemment et produire des effets divers. Ce n'est pas pour autant que la notion de signification lexicale n'est pas valide pour penser l'autonomie des unités lexicales.

On adoptera dans cette recherche, sans la justifier davantage, la position (iii) : l'autonomie des unités lexicales est de nature sémantique[4]. Autrement dit, on dira que l'autonomie des unités lexicales tient à leur signification lexicale, qui en conséquence doit présenter des caractéristiques de stabilité.

Dans le cadre adopté, l'une des façons d'aborder la notion de signification lexicale est de la penser par la notion de « stéréotype » proposée par Putnam comme étant l'une des dimensions possibles de la signification lexicale. Plus précisément, la notion de « stéréotype » constitue une « hypothèse sur la signification lexicale du point de vue de l'acquisition du langage[5] ». Autrement dit, si elle ne permet pas de répondre directement à la question : « qu'est-ce que la signification ? », elle donne des éléments de réponse à la question : « comment comprend-on les mots ? ». C'est en particulier pour cette raison qu'elle nous paraît pertinente pour notre étude des faits d'indexation.

Après avoir présenté la notion de stéréotype, nous montrerons la façon dont elle peut participer à la signification lexicale dans le cadre d'une approche morphologique qui étudie la construction du sens de certains types de mots : les mots construits.

---

[1] Marandin 1990, p. 290.
[2] *Id.*
[3] Marandin 1992a [présentation], p. 9 : « En tant qu'instance d'individuation, la signification lexicale est complexe et hétérogène : elle est externe aux unités lexicales (schèmes) et spécifique à chaque unité (stéréotypes sur les dimensions construites à partir des schèmes) ».
[4] Ou plutôt l'autonomie des unités lexicales est principalement de nature sémantique. Il y a, en effet, dans le cadre de ce modèle, deux autres facteurs d'individuation des unités lexicales : la forme phonologique et l'appartenance catégorielle (Milner 1989, p. 324) ; mais il reste que « le sens lexical est un facteur de différenciation *sine qua non* des atomes lexicaux », *Ibid.* p. 345.
[5] Marandin 1990, p. 286.

B - Approche de la notion de stéréotype

Si Putnam [1990a et 1990b] aborde le problème de la signification lexicale par le biais de la problématique « classique » de la stabilité du vocabulaire ordinaire, il oriente la question d'un point de vue particulier : comment la signification d'un mot demeure-t-elle identique alors que changent les croyances, les théories scientifiques, les usages ?

D'emblée est donc posée une distinction entre connaître le mot et connaître le fait que ce mot désigne. La viabilité de cette distinction est fondée sur ce que Putnam appelle la « division linguistique du travail » : l'idée est qu'il n'est pas nécessaire que tous les locuteurs connaissent les différences précises qui distinguent les « mots », « ils peuvent toujours se fier à des experts qui le feront à leur place[1] ». En ce sens, le langage est dit « coopératif », par opposition à la vision individualiste du langage proposée par les mentalistes[2]. Ce que souligne la notion de division linguistique du travail, c'est que le fait, la référence ou encore l'extension* n'a pas à faire partie de la signification pour qu'un locuteur puisse employer un mot. Ainsi est posée la possibilité d'une signification lexicale autonome, distincte des référents qu'une unité lexicale peut permettre de désigner : « Si communiquer la signification du mot "tigre" impliquait que l'on communique la totalité de la théorie scientifique acceptée, ou même la totalité de ce que je crois, à propos des tigres, ce serait une tâche impossible. C'est vrai que lorsque je dis à quelqu'un ce qu'est un tigre, "je lui dis simplement certaines phrases". [...] Le problème est donc bien quelles phrases ?[3] »

Le stéréotype correspond précisément à cet « ensemble de phrases », phrases qui décrivent une « théorie » extrêmement simplifiée à laquelle il n'est pas nécessaire de croire mais dont on doit *savoir* qu'elle est associée à un terme[4]. En ce sens, le stéréotype est « une idée conventionnelle, qui peut être fausse, sur un segment de la réalité [...] associée à un mot du langage naturel[5] ». Étant une convention, un « ensemble de croyances partagées[6] », le stéréotype définit moins un objet que la représentation de cet objet pour une communauté linguistique[7].

C'est pourquoi le stéréotype peut être faux : « Je peux référer à une espèce naturelle avec un terme qui est "chargé" d'une théorie dont on sait qu'elle n'est plus vraie de cette espèce, car tout le monde sait bien que mon intention est de référer à *l'espèce en question* et non de soutenir cette théorie.[8] »

La notion de stéréotype développée par Putnam opère une double césure par rapport aux théories traditionnelles de la signification lexicale et du lexique :

---

[1] Putnam 1990a, p. 54.
[2] Les mentalistes laissent entendre que « tout ce qui est nécessaire à l'usage du langage est emmagasiné dans chaque esprit individuel », Putnam 1990a, p. 57.
[3] Putnam 1990b, p. 299.
[4] *Ibid.*, p. 300.
[5] Marandin 1990, p. 285.
[6] Putnam 1990a, p. 96.
[7] Fradin et Marandin 1979, p. 65.
[8] Putnam 1990b, p. 300.

- traditionnellement, la signification lexicale se définit en termes de « conjonction de propriétés[1] ». Putnam montre, avec les exemples des citrons verts et des tigres à trois pattes, que cette conjonction de propriétés n'est le plus souvent que la description d'un « membre normal » d'une catégorie et que cette description est hétérogène (composants de nature linguistique et extra-linguistique) ;

- définir la signification lexicale en termes de conjonction de propriétés, c'est également, relève Putnam, adopter une conception unifiée et homogène du lexique : si une description analytique en termes de propriétés rend bien compte de mots comme *célibataire* (« homme qui ne s'est jamais marié »), elle n'est pas adaptée à tous les types d'unités lexicales[2]. Ce que les linguistes ont pu déduire de la thèse de Putnam, c'est que le lexique, loin d'être une « liste amorphe d'items », est hétérogène[3].

C'est parce que le lexique est doublement hétérogène, hétérogénéité d'unités (relevant d'une définition soit analytique soit stéréotypique) et hétérogénéité des composantes de la signification de ces unités (linguistiques et extra-linguistiques), que se justifie une théorie du lexique à même de distinguer des phénomènes de nature différente mais pouvant engendrer le même effet.

Une étude du lexique, dans cette approche, porte donc non plus sur les mots pris un à un mais sur des ensembles de mots regroupés en types.

Il importe donc, lorsque l'on s'intéresse à la signification lexicale comme ce qui constitue l'autonomie d'une unité lexicale, de considérer que :

- toutes les unités lexicales ne construisent pas la signification de la même façon, et qu'en ce sens il faut distinguer différents types de mots[4] ;

- la signification d'une unité lexicale ne correspond pas à « quelque chose d'unique et de bien circonscrit[5] » : la notion de stéréotype, l'une des dimensions de la signification lexicale, illustre cet aspect.

Notre recherche devra donc considérer que la définition du descripteur comme « substantif[6] » est trop large : il nous faudra essayer de déterminer de quels types de substantif relève le descripteur[7].

C - Exemple d'analyse de la signification lexicale vue sous l'angle du stéréotype

Pour montrer, sur un cas précis d'analyse de mots, comment le stéréotype peut être pris en compte dans la détermination de la signification lexicale, nous présentons

---

[1] Putnam 1990b, p. 292-296. Putnam critique le modèle de la définition analytique, qui peut prendre, de façon réductrice, la forme suivante : « *x* est un citron si *x* est jaune, avec un goût acidulé, une peau épaisse ».
[2] « On a soutenu que la théorie qui décrit correctement le comportement de peut-être trois cents mots décrivait correctement le comportement des termes généraux qui sont des dizaines de milliers ». Putnam 1990b, p. 294.
[3] Marandin 1990, p. 289.
[4] Putnam 1990b : « il y a différentes sortes de substantifs ».
[5] *Id.*
[6] Voir le discours normatif, présenté au § II.2 du chapitre I.
[7] Voir le chapitre V.

une analyse des mots *électrophone* et *tourne-disque* issue du cadre d'étude de la signification lexicale que propose D. Corbin pour ce qu'elle nomme les « mots construits[1] ».

Elle montre[2] que le sens des mots construits est étroitement associé à leur structure morphologique, « autrement dit qu'il existe un sens prédictible à partir de la façon dont le mot est construit[3] », ce qui n'exclut pas que des sens lexicaux différents puissent permettre de construire le même référent. Ainsi, si les deux mots *électrophone* et *tourne-disque* peuvent être utilisés pour renvoyer à la même catégorie d'objets, ils n'en suivent pas moins un parcours référentiel différent, lié aux propriétés différentes que chacun d'eux focalise : « *électrophone* décrit le mode de production du son, *tourne-disque* le fonctionnement de l'appareil[4] ». Cette focalisation différente porte sur des traits qui « reflètent *notre connaissance stéréotypique* des objets que ce mot dénomme. [...] En conséquence il faut admettre que la langue peut construire un sens en ne sélectionnant, dans le sens d'un mot, que des *traits stéréotypiques*[5] ».

Cet exemple issu d'un cadre d'analyse sémantique qui exploite le phénomène de stéréotypie montre comment la signification lexicale peut être ce qui permet d'assurer l'autonomie d'une unité lexicale. En effet, la signification d'un mot n'y est définie que par rapport à la langue elle-même : aucun élément extérieur – ni le concept, ni le référent – ne sont nécessaires à la détermination de la signification lexicale.

L'on voit en outre que ce qui détermine le rapprochement des termes, *électrophone* et *tourne-disque* par exemple : ce n'est pas leur signification mais leur référent. Se dessine ici un élément déterminant pour capter l'un des faits d'indexation : rappelons en effet que la synonymie documentaire se définit, dans le discours classique, contre la synonymie linguistique, en mettant en avant la ressemblance entre objets désignés par les descripteurs. Or la synonymie documentaire semble au contraire exploiter la notion de synonymie linguistique elle-même. Cependant cette exploitation ne peut apparaître que dans le cadre d'un modèle de fonctionnement de la langue qui distingue la signification lexicale (autonome, dont l'une des dimensions est stéréotypique) et la référence. Cette distinction n'étant pas opérée en indexation, les praticiens se trouvent conduits à « bricoler » de nouveaux concepts comme celui de synonymie documentaire.

On conclut sur ce point en mettant dans la perspective de l'indexation les propositions linguistiques ici présentées.

L'approche linguistique de la signification lexicale montre, nous semble-t-il, que l'indexation exploite, implicitement, par le biais du descripteur, une propriété des unités lexicales hors emploi : celle de leur autonomie, reposant en partie sur une représentation stéréotypique, qui présente des caractéristiques d'approximation. Nous dirons qu'en cela la notion de signification lexicale constitue l'un des

---

[1] « Ces mots sont construits par des opérations linguistiques, leur sens peut être calculé de façon proprement linguistique, indépendamment des catégories [référentielles] que les mots dénomment », Corbin et Temple 1994, p. 6.
[2] Ses études extrêmement précises ne peuvent être reprises ici. Le cadre général de l'analyse est présenté et argumenté dans Corbin 1987.
[3] Corbin et Temple 1994, p. 9.
[4] *Ibid.*, p. 10.
[5] *Ibid.*, p. 21. C'est nous qui soulignons.

fondements théoriques de l'indexation, dans le sens où elle fonde la possibilité de l'indexation : proposer à l'interprétation comme à l'utilisation des unités lexicales hors emploi. Une telle proposition peut paraître triviale et ne pas nécessiter une formulation en termes de fondements théoriques ; cependant nous avons vu que les approches normatives ne permettaient pas de poser la question de l'autonomie lexicale (c'est une évidence) et, encore moins, d'en proposer un traitement explicite dans le cadre documentaire.

En outre, il apparaît tout à fait nécessaire, pour comprendre la notion de synonymie en indexation, de faire apparaître une différence entre sens et référence. Nous précisons cet aspect ci-après.

II.2.2 - LA POSSIBILITÉ DE DÉSIGNATIONS MULTIPLES : LA QUESTION DE LA SYNONYMIE RÉFÉRENTIELLE

D'un point de vue linguistique, la question de la synonymie ne peut être posée que dans un cadre qui distingue les niveaux, notamment celui de la signification et celui de la référence, c'est-à-dire le niveau de la langue et celui du discours, ou encore celui du lexique et de la terminologie[1] : « Le lexique considère les mots, la terminologie considère les choses. Il n'existe pas d'équivalence d'un mot du lexique d'une langue à un mot du lexique d'une autre langue. Mais, si l'on se place dans la perspective de la terminologie, la même classe d'objets d'un univers donné peut avoir une étiquette dans une langue et une étiquette dans une autre langue ; dès lors, la traduction devient possible, fondée sur une synonymie référentielle. Si deux termes ont la même extension dans un univers donné, on peut les considérer comme équivalents, et les traduire l'un par l'autre.[2] »

Le modèle du lexique en indexation, modèle d'utilisation et modèle de fonctionnement, en posant une homogénéité des formes lexicales, ne peut distinguer ces niveaux ; il est donc amené à créer des concepts *ad hoc* comme celui de synonymie documentaire[3] ou de terme préférentiel[4] qui assimilent les deux types de fonctionnement du descripteur :

- son autonomie, qui correspond au niveau du « contenu » en indexation, au niveau de la signification en linguistique, au niveau du lexique dans le cadre logico-sémantique ;

- sa « dépendance », qui correspond, en indexation, à la relation qu'il établit avec les documents, à la référence en linguistique, au niveau de la terminologie dans le cadre logico-sémantique.

---

[1] Dans le cadre du modèle logico-sémantique que propose Le Guern [1989, p. 340] : « On peut dire que le lexique concerne les mots indépendamment des choses, alors que dans la terminologie, les mots sont liés aux choses. Mais d'un côté et de l'autre, ce ne sont pas les mêmes "mots". Ils ont bien l'air d'être les mêmes, et beaucoup de gens s'y trompent, mais l'objet "mot" pertinent pour le lexique est une réalité totalement distincte de l'objet "mot" qui appartient à la terminologie ».

[2] Le Guern 1989, p. 342.

[3] Cf. Maniez 1987, Van Slype 1987, Chaumier 1988.

[4] Cf. la définition normative du descripteur [norme AFNOR Z 47-100 (1981)] : « Mot ou groupe de mots retenus dans un thésaurus et *choisis parmi un ensemble de termes équivalents* pour représenter sans ambiguïté une notion contenue dans un document ou une demande de recherche documentaire ». C'est nous qui soulignons.

Dans le discours classique, le descripteur est décrit comme étant, en lui-même, de par sa « nature », un synonyme, un nom de classe d'équivalence. La signification lexicale, qui n'est pas explicitement posée, y est donc confondue avec la référence, selon un mécanisme qu'exprime ainsi Marandin sur l'exemple du couple *tête / caboche* : « Au regard d'une conception qui assimile signification lexicale et description de l'extension [ou référence], *tête* et *caboche* sont synonymes ; *tête* et *caboche* ne sont pas synonymes si on admet que la description de l'extension n'épuise pas la signification lexicale.[1] »

Pour capter le fonctionnement spécifique du descripteur en tant que relais textuel, c'est-à-dire comme forme lexicale susceptible de convenir à plusieurs documents, il importe donc de distinguer les niveaux de façon à pouvoir préciser celui auquel se greffe la synonymie : niveau de la langue ou niveau du discours.

On peut montrer que la notion de « synonymie » n'est pas statiquement attachée à une unité lexicale et qu'en cela la synonymie référentielle relève du discours ; sur ce point, nous reprenons l'analyse que propose Franckel à partir des deux verbes *manger* et *bouffer* :

- dans les contextes de type *On a bien mangé / On a bien bouffé*, on peut considérer qu'il y a synonymie ;

- la substitution « manger/bouffer » est moins nette avec l'exemple : *Ça se laisse manger* (paraphrasé par *C'est mangeable*) ;

- dans le cas *Il se laisse bouffer par son travail*, c'est moins « manger » qui apparaît comme un bon candidat-synonyme que les verbes « déborder » ou « accaparer ».

Cet exemple montre que si, dans certains contextes, deux termes paraissent substituables, ils n'en sont pas pour autant intrinsèquement « équivalents ».

Il y a là, comme le relève Marandin, une « illusion » dans le sens où le rapprochement de deux unités n'est « l'indice d'aucune ressemblance sémantique[2] ».

On remarquera que le discours normatif notifie cet aspect dans la définition qu'il donne de la synonymie documentaire[3], mais qu'il en fait une caractéristique distinctive du descripteur par rapport au mot de la langue. Or, comme nous le montrent les propositions linguistiques, il s'agit moins d'une spécificité du descripteur que d'une propriété de langue, et plus précisément, une propriété liée à la signification lexicale.

En effet, « multidimensionnelle » ou encore hétérogène, la signification lexicale peut être à ce titre à l'origine de manifestations diverses, pour lesquelles Marandin propose la formulation suivante : « *Il y a de la synonymie* (dans le lexique virtuel),

---

[1] Marandin 1992a, p. 40.
[2] *Ibid.*, p. 46.
[3] Cf. Chaumier 1978, p. 33 : « La notion de synonymie est utilisée de façon extensive dans les thésaurus sous la forme de la synonymie documentaire afin de regrouper sous un seul descripteur plusieurs termes considérés comme voisins, *bien que de signification sémantique différente* », c'est nous qui soulignons.

il y a des effets de synonymie contextuelle (dans des énoncés dans certain contexte) et *il n'y a pas de synonymes* (dans les énoncés actuels).[1] »

En effet, comme Le Guern, Marandin pose qu'il ne s'agit pas des « mêmes » mots selon les dimensions considérées. Marandin distingue en effet plusieurs types d'individus linguistiques, notamment les occurrences des unités lexicales (qui, dans certaines contextes, peuvent avoir des effets synonymiques) et les unités lexicales « hors emploi » (qui, à un certain niveau d'abstraction, peuvent ne pas être distinguées[2]).

Cette distinction, entre type (unité hors emploi) et occurrence (unité en contexte), est au cœur de l'explication de la synonymie documentaire que propose Le Guern[3], à la différence près que les présupposés de la démonstration relèvent ici autant de la logique que de la linguistique. Dans son cadre, l'ensemble des unités hors emploi constitue le « lexique », dont les éléments n'ont qu'une intension* (pas d'extension, de référence) : il n'y a pas, à ce niveau, de synonymie, de traduction possible. En revanche, les unités en emploi (en discours) relèvent d'un niveau, celui de la « terminologie », où les traits référentiels permettent de concevoir l'identité référentielle et par suite la synonymie référentielle (ou documentaire[4]) ; c'est là où Marandin parle d'« effets synonymiques ». Pour Le Guern, le descripteur, en tant qu'il réfère d'une part et qu'il condense un ensemble de synonymes référentiels d'autre part, est donc un « terme », une unité de discours[5].

Or, tout le pari de l'indexation consiste à faire fonctionner cette unité de discours qu'est le descripteur sous la forme d'une unité de langue (autonome), tentant ainsi de conjoindre deux propriétés de langue différentes : en effet,

- en tant qu'élément d'une liste (liste d'un langage documentaire ou liste d'une formule d'indexation), le descripteur fonctionne comme une unité lexicale hors emploi : à ce titre, on peut dire que le descripteur est doté d'une autonomie lexicale qui vient de sa signification. Nous sommes là au niveau de la langue ;

- en tant que forme lexicale lue, interprétée et utilisée dans le contexte des documents auxquels elle est affectée, le descripteur fonctionne comme un synonyme référentiel : à ce titre, il peut permettre de rapprocher des « objets », des documents différents. Nous sommes là au niveau du discours, de l'emploi des formes lexicales.

Alors que l'approche linguistique, parce qu'elle distingue les niveaux, distingue aussi les unités, unité de langue et unité de discours, l'approche documentaire, elle, tendant à assimiler les niveaux, assimile aussi des propriétés différentes sous une même forme linguistique. En cela, si les concepts linguistiques de signification lexicale et de synonymie référentielle peuvent être vus comme des fondements théoriques de l'indexation, dans le sens où ils fondent, du point de vue d'une théorie linguistique, la pratique de l'indexation telle qu'elle s'exerce, ils ne peuvent

---

[1] Marandin 1992a, p. 53.
[2] C'est le cas particulier de certains déverbaux : triche/tricherie. Marandin 1992a, p. 53.
[3] Le Guern 1984 et 1989.
[4] « Deux descripteurs sont synonymes s'ils ont la même référence ; il ne s'agit donc pas, dans une perspective documentaire, de synonymie lexicale, mais de synonymie référentielle », Le Guern 1984, p. 167.
[5] Nous reviendrons largement sur ce point dans le chapitre V.

l'être que dans le cadre d'un modèle d'utilisation du lexique : dans le modèle théorique de la langue, ces deux propriétés restent incompatibles (une forme lexicale n'est pas la « même » selon qu'elle est vue hors ou en emploi). C'est sous cet angle que nous proposons, dans le paragraphe III, un modèle d'utilisation du lexique en indexation qui, tout en étant fondé sur des propriétés linguistiques, laisse de quoi penser la torsion que réalise l'indexation.

II.3 - Conclusion et résultats intermédiaires

Une lecture du discours normatif sur l'indexation permet de dégager cinq caractéristiques du descripteur et parallèlement des représentations de la langue propres au modèle objectiviste (I) :

- caractéristique 1 : le descripteur comme expression linguistique d'un concept préexistant, stable, unitaire (« simple ») ;
- caractéristique 2 : le descripteur comme expression linguistique stable ;
- caractéristique 3 : le descripteur comme condensateur textuel ;
- caractéristique 4 : le descripteur comme expression linguistique autonome ;
- caractéristique 5 : le descripteur comme relais textuel.

Le déplacement du modèle de fonctionnement de la langue, d'un modèle objectiviste à un modèle linguistique, montre que (II) :

- les caractéristiques 1, 2 et 3 du descripteur ne peuvent être appréhendées dans un cadre linguistique. En revanche, la fonction 1 de l'indexation à la base de la formulation de ces caractéristiques peut, elle, être analysée dans le cadre d'un modèle linguistique en termes d'effet d'interprétation : effet obtenu sur la base des descripteurs qui fonctionneraient comme thèmes de discours.

  Cette hypothèse du descripteur comme thème de discours conduit à chercher comment l'indexation peut permettre à un utilisateur de construire ces thèmes de discours : nous faisons l'hypothèse que l'indexation met en œuvre une stratégie de regroupement et d'exposition des documents qui permette de telles constructions (*infra*, chapitre IV) ;

- les caractéristiques 4 et 5 du descripteur, qui découlent de la seconde fonction assignée à l'indexation, rencontrent, elles, des propriétés de langue : le descripteur peut être utilisé comme un accès autonome parce que, en tant qu'unité lexicale hors emploi, il est pourvu d'une signification lexicale ; le descripteur peut être utilisé comme synonyme parce que, en tant qu'unité de discours, il peut renvoyer à des objets référentiellement différents.

Ces aspects de la signification lexicale et de la synonymie référentielle en jeu dans l'indexation *via* son résultat (le descripteur), pourraient constituer directement des fondements théoriques s'ils étaient utilisés, dans l'indexation, d'une façon qui les maintienne distincts. Or, l'indexation attribue à une même forme lexicale des propriétés distinctes, « incompatibles ». C'est ainsi qu'il devient nécessaire d'essayer de concevoir un modèle d'utilisation du lexique qui pense ces différents niveaux et les articule.

III - Reformulation du modèle d'utilisation du lexique en indexation

Nous avons montré en I que le modèle d'utilisation du lexique en indexation reposait sur des représentations de la langue non valides du point de vue de la théorie linguistique et non complètes d'un point de vue descriptif. Nous avons vu en II qu'un modèle linguistique pouvait proposer des explications du fonctionnement des mots en indexation. Si nous adoptons les représentations linguistiques de la langue que nous avons présentées en II, que reste-t-il de l'indexation et du descripteur tels qu'ils ont été approchés en I par le biais des normes ? À quelles définitions de l'indexation (III.1) et du descripteur (III.2) arrivons-nous ?

III.1 - L'indexation dans le cadre d'une approche linguistique du lexique

Dans le nouveau cadre de fonctionnement du lexique adopté, la définition normative de l'indexation n'est plus valide.

On se souvient que l'indexation est définie dans la norme comme un processus réalisé en deux phases : une phase d'analyse conceptuelle (analyse de contenu) et une phase de représentation linguistique (traduction). Ce processus en deux phases semble devoir être mis en cause (III.1.1) ; par suite, cette remise en cause conduit à mettre à distance le rôle des mots en indexation (III.1.2).

III.1.1 - MISE EN CAUSE DU PROCESSUS EN DEUX PHASES

Il apparaît que la description de l'indexation en deux phases est artificielle : la reconnaissance de concepts d'une part et leur codage par des descripteurs d'autre part ne valent que dans le cadre d'un modèle objectiviste de la langue. Hors de ce modèle, le processus en deux phases ne peut être vu que comme une décomposition méthodologique établie *a posteriori* à des fins didactiques, ou du moins supposées didactiques.

La faillite de la description d'un processus en deux phases, si elle ressort d'une approche théorique au terme d'une confrontation entre modèles de langue, a pu être mise en évidence sur la base d'expérimentations menées dans le cadre des sciences cognitives. Ainsi Sylvie Bruxelles [1991], qui a étudié, d'un point de vue linguistique et psychologique, les codages réalisés par plus de 150 sujets à partir de deux nomenclatures[1], remarque-t-elle que :

(i) le codage, la classification, ne relève pas de la traduction (mot à mot, concept à concept, ou encore concept à mot) mais de l'intertextualité, de la confrontation de deux textes. En effet, en montrant les contraintes linguistiques et cognitives exercées par les nomenclatures sur le choix des postes de classement, Bruxelles note que l'instrument d'indexation constitue le

[1] Bruxelles 1991, p. 171-186 : l'expérimentation porte sur des documents juridiques (40 assignations et requêtes issues de juridictions) indexés à l'aide de deux nomenclatures (plus précisément à l'aide de deux versions de la *Nomenclature des affaires civiles*, celle de 1981 et celle de 1988) par trois groupes d'individus : 92 greffiers, 42 élèves de l'École nationale de la Magistrature, 30 étudiants non juristes.

texte d'arrivée qui permet de lire le texte-source (le document à indexer). Outre que le codeur, contrairement au traducteur, dispose d'emblée des deux textes (source et cible), l'orientation de leur activité est inverse : la traduction opère de la source à la cible alors que l'indexation lit la source en fonction de la cible ;

(ii) en ce sens, le codage ne suppose pas une analyse de contenu préalable, entendue comme une « décomposition analytique des propriétés des objets manipulés. Les associations sémantiques se forgent plutôt par agrégation à partir de foyers-repères où se nouent des « reconnaissances[1] ». En ce sens, les nomenclatures apparaissent comme des « principes de construction de l'information[2] » et non comme de simples outils de transmission d'une information préexistante. C'est pourquoi le document, le texte classé, apparaît comme un construit, ou un reconstruit[3].

Ces résultats d'analyse d'activités de codage, succinctement rapportés ici, confirment certaines de nos conclusions, établies sur un plan plus formel :

- le processus de l'indexation peut être dégagé des notions d'analyse de contenu et de traduction (point (ii)) ;

- sur le plan de l'utilisation des mots (de l'attribution des descripteurs), l'indexation semble se réaliser en une seule phase (point (i)), la fonction de fourniture d'accès supplantant celle de représentation du contenu ;

- l'indexation apparaît comme une opération de construction du document (point (ii)).

L'utilisation des mots en indexation ne semble donc pas relever de la traduction, ni de l'analyse de contenu, mais plutôt de la construction de l'information qui se ferait *via* l'établissement d'accès aux documents et *via* la construction des documents eux-mêmes. D'autres mécanismes langagiers sont alors à postuler. Pour ce qui est de l'indexation « contrôlée », Bruxelles suggère que les relations d'intertextualité sont à prendre en compte dans le processus de l'indexation. Nous ferons l'hypothèse plus large que le principe de la confrontation textuelle est à l'œuvre dans tout type d'indexation (chapitre IV).

III.1.2 - MISE À DISTANCE DES MOTS EN INDEXATION

La mise en cause de l'indexation comme traduction d'un contenu préexistant induit des conséquences à différents niveaux, qui conduisent à mettre à distance le rôle des mots en indexation.

A - Conséquences sur la fonction de l'indexeur

L'indexeur ne peut plus être vu comme un traducteur, pris dans un rapport paradoxal avec le langage, considéré sous la double facette du « coupable » et du « rédempteur ». On retrouve ici la fin du mythe du « médiateur », que l'analyse

---

[1] Bruxelles 1991, p. 182.
[2] *Id.*
[3] *Ibid.*, p. 183 : Bruxelles note que les codeurs « mettent en œuvre des procédures de reconstruction d'objets, dominées par des effets de contexte ».

d'autres activités de diffusion de connaissances, comme la vulgarisation scientifique, a pu mettre en valeur[1] ; ainsi celle menée par Jacobi : « La science ne serait pas comprise avant tout parce qu'elle se parle dans une langue ésotérique. [...] Cette idéologie impose la figure du médiateur qui, par sa compétence de traduction, parvient à rétablir la communication. [...] Dans cette perspective, le langage apparaît sous le double visage du coupable et du rédempteur.[2] »

Sur ce point, il apparaît que le travail de l'indexeur doit porter moins sur une manipulation de mots (trouver les « bons » mots qui permettront aux utilisateurs de communiquer avec des auteurs) que sur une mise en contexte des mots susceptible de permettre cette communication.

B - Conséquences sur l'approche de la langue

Le lexique, l'utilisation de mots en indexation, n'engage pas forcément, comme l'indique le modèle linguistique, une approche strictement lexicaliste de la langue ; il peut, au contraire, ouvrir la voie au discours, à l'analyse des discours. C'est une tout autre appréhension du lexique qui peut alors être adoptée : « Aussi faut-il voir dans le lexique moins une donnée contraignante, dont l'emploi serait soumis au seul principe d'adéquation référentielle, qu'un ensemble de dispositifs extrêmement malléables, continuellement travaillés et retravaillés dans et par le discours.[3] »

Si elle s'exprime prioritairement par les mots, l'indexation n'est donc pas uniquement concernée par eux. Nous aurons donc à dégager la dimension discursive de l'indexation.

C - Conséquences sur l'approche du langage documentaire

La notion de langage documentaire est amenée, elle aussi, à être repensée. Elle se donne traditionnellement pour un métalangage, substituant à des noms d'autres noms (traduction mot_1 –> mot_2 dans le schéma classique de l'indexation). Or, dans le cadre du modèle de fonctionnement de la langue adopté, « le métalangage est une illusion logique : il n'y a pas de noms de noms, mais seulement des noms de gestes, d'événements, de choses[4] » ; la notion de métalangage ne tient que si l'on ne distingue pas les niveaux, ici celui des mots et celui des choses[5]. L'enjeu du langage documentaire n'est donc plus celui d'être un outil de représentation du contenu mais plutôt celui d'être un outil de construction du contenu (Bruxelles 1991, *supra*).

Il apparaît que le déplacement du modèle de fonctionnement du lexique tel qu'on le propose opère un renversement des données : ce n'est plus, en indexation, la représentation du contenu d'un document qui permet l'accès à l'information d'un fonds documentaire, c'est plutôt la fourniture d'accès aux documents qui permet de construire une représentation du contenu d'un document.

---

[1] On revient sur ce point dans le chapitre IV, § I.3.
[2] Jacobi 1987, p. 26.
[3] Apothéloz et Reichler-Béguelin 1995, p. 241.
[4] Berrendonner 1981, p. 132.
[5] *Id.*, note 19 : « Je me refuse d'admettre que les objets situés à deux niveaux contigus soient des objets du même ordre : pour moi, la référence relie toujours une chose et un nom, et la chose ne saurait être un nom ».

L'enjeu de l'indexation se pose alors moins dans le cadre du lexique (niveau de la signification lexicale, du contenu) que dans celui de la référence (niveau des référents et des discours). L'indexation rejoint par là des problématiques plus générales, communes à d'autres pratiques : « Le problème n'est donc plus de se demander comment l'information[1] est transmise ou comment des états du monde sont représentés de façon adéquate, mais de se demander comment les activités humaines, cognitives et linguistiques, structurent et donnent un sens au monde.[2] »

Dans ce cadre, l'indexation n'a plus à faire se rencontrer des « mots » mais plutôt des « mondes », ceux des auteurs et ceux des utilisateurs, et ceci par le biais de mots, les descripteurs. L'approche du descripteur en indexation est donc nécessairement au moins double : elle se mène à la fois du côté du lexique et du côté de la référence.

III.2 - Le descripteur dans le cadre d'une approche linguistique du lexique

Notre approche de l'indexation nous conduit à considérer le descripteur comme un « accès documentaire », c'est-à-dire comme une forme lexicale susceptible de conjoindre des propriétés linguistiquement incompatibles :

- pour constituer un accès autonome, le descripteur doit être appréhendé comme une unité lexicale hors emploi, pourvue d'une autonomie : sa signification lexicale, qui révèle une certaine stabilité (stabilité sémantique de nature stéréotypique, largement sous-déterminée) ;

- pour constituer un accès « multiple » (à plusieurs documents), le descripteur doit être une unité de discours pour fonctionner, aux yeux des utilisateurs, comme un thème de discours susceptible de construire des effets de synonymie référentielle.

Sur un plan linguistique, les deux niveaux de langue et de discours sont tenus pour distincts et l'individu de langue n'est pas le même que l'individu de discours. Or, l'indexation a besoin d'utiliser les mots sur les deux plans en même temps : dans le cadre de quel modèle d'utilisation du lexique en indexation peut-on rendre compte de cette dualité du descripteur, à la fois type ET occurrence ?

On se propose, à la suite de Michel Le Guern et au travers de sa propre analyse, d'appréhender la relation type/occurrence dans le cadre proposé par Peirce : Le Guern propose en effet une lecture de Peirce qui permet de voir comment le descripteur peut fonctionner à un double niveau de langue et de discours[3].

---

[1] Le terme « information » est à prendre ici dans l'acception que lui donnent les sciences cognitives.

[2] Dubois et Mondada 1995, p. 276.

[3] La démarche de Le Guern [1991a] s'inscrit dans le cadre d'une entreprise de « désintrication » des niveaux, non distingués dans le discours classique en raison de l'absence de référentiel théorique : « Il est fâcheux que l'on ait appelé les descripteurs des « mots-clés », ce qui laisse croire que ce sont des mots, des unités lexicales. Or, dans les pratiques documentaires les plus courantes, l'indexation vise les objets, les référents, et non les signifiés. Cette distinction revêt d'autant plus d'importance qu'elle se heurte plus fortement à la conception naïve – généralement répandue chez les utilisateurs de la langue

Il ne s'agit pas, pour nous, d'exposer le détail de cette approche : nous nous contenterons, d'une part, d'indiquer quelques notions générales pour situer la problématique de Peirce et nous présenterons, d'autre part, certains des concepts de Peirce dégagés par Michel Le Guern comme pertinents en matière d'explicitation du mécanisme d'indexation.

III.2.1 - BRÈVE PRÉSENTATION DU MODÈLE DE PEIRCE

Rappelons, pour commencer, que la théorie sémiotique de Peirce ne s'appuie pas sur des présupposés de nature linguistique ; relevant d'une réflexion d'ordre phénoménologique, elle explore plutôt la logique des relations[1]. Les travaux de Peirce sont néanmoins considérés comme étant à l'origine de la distinction, formalisée ensuite par Morris, entre les trois « niveaux » de l'analyse linguistique, syntaxe, sémantique et pragmatique[2].

Les principales notions que manipule Peirce sont le signe et la sémiosis, dont la problématique peut s'énoncer de la façon suivante : « Toute chose, tout phénomène, aussi complexe soit-il, peut être considéré comme *signe* dès qu'il entre dans un processus sémiotique, c'est-à-dire dès qu'un interprète le réfère à autre chose.[3] » « La *"sémiosis"* ou la production de la signification est un processus triadique qui met en relation un representamen, un objet et un interprétant.[4] »

De façon schématique, on entend par :
- representamen : une chose qui représente une autre chose ;
- objet : une entité physique ou mentale ;
- interprétant : une action de médiation entre le representamen et l'objet[5].

Chacun de ces trois éléments se subdivise encore lui-même en trois catégories :

- selon l'ordre dont il relève[6], un representamen sera un qualisigne (priméité), un sinsigne (secondéité) ou un légisigne (tiercéité) ;

- selon le type de relation que l'objet entretient avec le representamen, l'objet sera un icône (relation de similarité), un indice (relation de contiguïté contextuelle) ou un symbole (relation arbitraire, provenant d'une règle, d'une loi, etc.) ;

---

qui ne sont ni linguistes ni logiciens – des relations entre les mots et les choses. Cette conception, d'après laquelle les mots, en tant qu'unités lexicales et préalablement à toute insertion dans le discours – ou tout au moins certains d'entre eux –, désigneraient directement les choses, peut être appelée l'illusion du substantif. Cette illusion remonte au moins à Aristote, et il n'était sans doute pas possible d'y résister tant qu'on n'apercevait pas de manière nette l'opposition langue / parole », Le Guern 1991a, p. 23.

[1] Voir Deledalle *in* Peirce 1978, p. 212 : « La sémiotique est, selon Peirce, un autre nom de la logique : "la doctrine *quasi* nécessaire ou formelle des signes" ».

[2] On peut trouver une présentation générale de l'approche de Peirce dans Everaert-Desmedt 1990. Nous nous sommes en grand partie appuyée sur cette présentation dans le paragraphe qui suit.

[3] Everaert-Desmedt 1990, p. 25.

[4] *Ibid.*, p. 26

[5] Peirce précise que « l'interprétant n'est pas l'interprète mais le moyen que celui-ci utilise pour effectuer son interprétation », *in* Everaert-Desmedt 1990, p. 40.

[6] Peirce distingue trois ordres : la priméité (le possible), la secondéité (le réel) et la tiercéité (la loi), qui constituent les principes de subdivision en catégories à la fois dans le representamen, l'objet et l'interprétant, voir Everaert-Desmedt 1990, p. 48 notamment.

- selon le type de règle qui renvoie le representamen à son objet, l'interprétant sera un rhème (règle reposant sur les caractères de l'objet seulement), un dicisigne (règle reposant sur l'existence de l'objet) ou un argument (règle exploitant l'objet en tant que signe).

Pour qu'il y ait signe, il faut, comme nous l'avons vu précédemment, que s'établisse un « processus triadique », c'est-à-dire une relation entre un representamen, un objet et un interprétant. En fonction des trois trichotomies ci-dessus présentées, on devrait avoir 27 types de signe possibles ; mais, ces trichotomies étant fondées sur des « ordres », exprimant une hiérarchie, certaines des combinatoires possibles ne sont pas valides. C'est ainsi qu'on retient généralement dix modes de fonctionnement de la signification (ou « signes »)[1].

III.2.2 - APPROCHE DU DESCRIPTEUR DANS LE MODÈLE DE PEIRCE

C'est à partir du cadre posé par Peirce que Michel Le Guern[2] propose de faire « voir » la différence et la ressemblance entre mot de la langue et descripteur[3] : il situe le point de distinction sur la façon dont chacun de ces « signes » gère la relation type/occurrence, c'est-à-dire la relation representamen/objet ou encore mot/chose.

Ainsi Michel Le Guern explique-t-il que le descripteur peut s'analyser comme un « légisigne indiciaire rhématique », tandis que le mot de la langue est, lui, un « légisigne symbolique rhématique ». Les deux types de signe sont des « légisignes[4] », dont la propriété est de ne « pouvoir agir qu'en se matérialisant dans des sinsignes qui constituent des répliques ». En effet, le légisigne est un « type général », accessible uniquement par le biais de ses occurrences : « En soi, un signe est soit une apparence, ce que j'appelle un *qualisigne*, soit un objet ou événement individuel, ce que j'appelle un *sinsigne* (la syllabe *sin* étant la première syllabe de *sem*el, *sim*ul, *sin*gulier, etc.), soit un type général, ce que j'appelle un *légisigne*. Comme nous employons le terme "mot" dans la plupart des cas, quand nous disons que "le" est un "mot", que "un" est un autre mot, un "mot" est un légisigne. Mais quand nous disons d'une page d'un livre qu'elle a deux-cent-cinquante "mots" dont vingt sont des "le", le "mot" est un sinsigne. Un sinsigne qui renferme ainsi un légisigne, je l'appelle une "réplique" du légisigne.[5] »

---

[1] Ces dix signes sont d'après Everaert-Desmedt 1990, p. 94 :

qualisigne iconique rhématique	légisigne indiciaire rhématique
sinsigne iconique rhématique	légisigne indiciaire dicent
sinsigne indiciaire rhématique	légisigne symbolique rhématique
sinsigne indiciaire dicent	légisigne symbolique dicent
légisigne iconique rhématique	légisigne symbolique argumental

[2] Le Guern 1984, 1989, 1991a.

[3] La démonstration menée par Le Guern a pour objectif de déplacer l'appréhension du descripteur d'une sémantique lexicale à une sémantique discursive ; dans cette perspective, ce qui est opposé, c'est le mot du lexique (appelé, dans ce cadre, « mot de la langue ») et le mot du discours (en l'occurrence le descripteur). Si l'on veut bien s'abstraire des différences de terminologie, on conviendra que les vues exprimées ici convergent avec celles précédemment exposées dans le § II.

[4] « Le légisigne est un signe dont le fondement est une loi. Une loi est établie *a priori*, par convention, décision arbitraire ; ou *a posteriori*, par habitude ». C'est pourquoi il a une « identité bien déterminée », Everaert-Desmedt 1990, p. 51.

[5] Peirce, cité *in* Deledalle 1990, p. 85.

Mais, remarque Michel Le Guern, selon qu'il est « indiciaire » ou « symbolique », le légisigne n'établit pas la même relation entre type et occurrence[1] : en tant qu'indice, le descripteur ne fait que *désigner* son objet, alors que le mot de la langue, étant symbole, signifie son objet par l'intermédiaire d'un *interprétant*. En effet, le rôle de l'indice est d'assurer la référence et de ne faire que cela : « sa fonction est pragmatique et non sémantique[2] » ; par opposition, le symbole établit une relation indirecte avec son objet, par l'intermédiaire de l'interprétant : « Un symbole est un signe qui perdrait le caractère qui en fait un signe s'il n'y avait pas d'interprétant. Exemple : tout discours qui signifie ce qu'il signifie par le seul fait que l'on comprenne qu'il a cette signification.[3] »

En simplifiant, on pourrait dire que, dans le cas du légisigne indiciaire rhématique, la relation type/occurrence est de type référentiel ; elle est de nature « sémantique » dans le cas du légisigne symbolique rhématique.

Cependant, comme le précise Le Guern, « le mot de la langue est également l'interprétant du descripteur[4] » : en effet, l'indice a simplement pour objet de montrer un objet, « il appartient au symbole d'en parler[5] ».

Autrement dit, en tant qu'indice, le descripteur est une unité de discours : en ce sens il peut, *via* ses occurrences, désigner différents objets singuliers, et permettre de constituer, de texte en texte, un thème de discours. En tant qu'indice toujours, le descripteur entretient une relation avec l'unité lexicale hors emploi (son interprétant) : la signification lexicale que se voit alors attribué le descripteur, notamment par sa sous-détermination référentielle, permet de le doter d'une certaine stabilité sémantique, qui ne préjuge en rien de son instabilité référentielle.

C'est par l'établissement de ces deux fonctions du descripteur en tant qu'indice (« montrer » et « dire ») que Michel Le Guern propose de voir le descripteur comme un nom propre[6] tel qu'il est entendu dans la typologie de Peirce[7]. Perçu comme légisigne indiciaire rhématique[8], le nom propre constitue pour Peirce un véritable paradoxe, dans le sens où ce qu'il représente ne se réduit jamais à ce qui en est dit : « Affronté à la nécessité de désigner des événements singuliers par le moyen des termes ayant une signification générale, le langage trouve, dans le nom propre, un moyen de dépasser embrayeurs et descriptions : tandis que les premiers montrent sans rien dire (index purs) et que les secondes disent sans montrer (symboles iconiques purs), le nom propre en tant qu'articulation symbolico-iconico-

---

[1] Le Guern 1984, p. 165.
[2] Everaert-Desmedt 1990, p. 64.
[3] Peirce cité *in* Deledalle 1990, p. 86.
[4] Le Guern 1984, p. 165.
[5] Deledalle *in* Peirce 1978, p. 235.
[6] « Si le descripteur est le signe de ses occurrences dans le corpus, ce n'est pas comme s'il en était en quelque sorte la reproduction photographique. Il en est plutôt le "nom propre" ». Le Guern 1984, p. 166.
[7] Voir Thibaud 1989, p. 381 : « Une expression est un nom propre si et seulement si il est possible de l'introduire comme index d'un objet individuel, de telle sorte que ce nom peut être utilisé comme symbole dans des situations différentes de celle où l'objet est présent et indexiquement reconnu ».
[8] Par opposition au nom commun, compris comme un légisigne symbolique rhématique, qui, en tant que symbole, ne peut en lui-même « identifier les choses », Deledalle *in* Peirce 1978, p. 165.

indexique est ce qui relie un dire à une monstration. On peut donc dire qu'il est la voie d'accès privilégiée à l'individuel.[1] »

En suivant la proposition de Le Guern, le descripteur, appréhendé comme un nom propre au sens que Peirce attribue à ce mot, apparaît donc comme une unité qui « désigne un objet que n'épuise aucune description[2] » : la signification lexicale du descripteur n'épuise pas, en effet, sa référence.

Ce modèle du descripteur comme nom propre de ses occurrences, proposé par Michel Le Guern sur la base de la théorie peircienne du signe, permet, nous semble-t-il, de comprendre comment l'indexation peut articuler signification lexicale et synonymie référentielle et conjoindre ainsi la stabilité de la signification avec l'instabilité de désignation.

En outre, ce modèle dégage la notion de nom propre, que nous avons par ailleurs déjà vu apparaître, et précise l'usage que l'on peut en faire dans le cadre d'un modèle d'utilisation de la langue : si le descripteur fonctionne comme un nom propre, il s'agit là d'une caractéristique de fonctionnement et non d'une propriété de « nature ». Le descripteur n'a pas à être un nom propre sur un plan linguistique mais doit se trouver doté de ses particularités de fonctionnement, notamment celle de désigner des objets individuels. Cette propriété lui vient de sa catégorie grammaticale : le nom propre est un groupe nominal. Le descripteur devra donc pouvoir être un groupe nominal, en discours, en même temps qu'une unité lexicale, hors emploi : c'est sur cette base que nous développerons notre approche du descripteur dans le chapitre V.

Enfin, cette approche du descripteur « fonctionnant comme un nom propre » met au jour l'enjeu de l'indexation : mettre à disposition des utilisateurs des procédés de désignation stable. Ces procédés ne relèvent pas uniquement des mots eux-mêmes, du lexique : nous aurons à introduire la dimension du discours pour comprendre comment l'indexation peut réaliser un tel pari.

IV - Conclusion du chapitre

Pour peu apparente qu'elle soit dans le discours normatif, la dimension lexicale de l'indexation nous paraît essentielle à faire apparaître. On peut en effet, par elle, d'une part, distinguer ce qui dans l'indexation relève des effets d'utilisation de la langue et ce qui relève des propriétés de la langue elle-même et, d'autre part, approcher la finalité de l'indexation dans des termes qui ne soient pas circulaires.

Sur ces deux points, ce chapitre a permis d'obtenir les résultats suivants :

(i) Distinction des faits et des effets

- Analyse des effets et conséquences sur la poursuite de la recherche

---

[1] Thibaud 1989, p. 386.
[2] *Ibid.*, p. 384.

Nous sont apparus comme pouvant être des effets d'interprétation les aspects de l'indexation relevant de l'analyse et de la représentation du contenu. Cette approche en termes d'effets nous conduit à poser l'hypothèse que le descripteur doit pouvoir fonctionner, aux yeux d'un utilisateur, comme un thème de discours. Une telle hypothèse nécessite de considérer l'indexation comme un processus particulier d'organisation de textes (notion d'interdiscours en indexation). D'où la nécessité de poursuivre la recherche en abordant l'indexation sous l'angle du discours[1]. L'analyse du thème de discours ne relève pas, en effet, d'une sémantique lexicale, mais d'une sémantique discursive.

- Analyse des faits et première approche des fondements théoriques de l'indexation

En proposant à l'utilisation des « mots » isolés, l'indexation nous est apparue comme exploitant une propriété des unités lexicales hors emploi : celle de leur autonomie lexicale, que l'on a proposé d'approcher sous l'angle de la signification lexicale. Parallèlement, nous avons relevé que l'indexation exploitait aussi le fonctionnement synonymique des unités lexicales en discours : ce n'est qu'en discours qu'une unité lexicale peut désigner des « objets » différents.

Si l'indexation exploite des propriétés spécifiques aux unités lexicales de la linguistique, elle les exploite d'une façon qui les rend incompatibles au regard d'une théorie de la langue. Si les propriétés d'autonomie lexicale et de synonymie référentielle constituent des fondements théoriques de l'indexation, ce ne peut être que dans le cadre d'un modèle d'utilisation de la langue qui permet de les articuler de façon non contradictoire. C'est ainsi que nous avons proposé de commencer à bâtir ce modèle d'utilisation : le descripteur peut être considéré comme le nom propre de ses occurrences dans un corpus. Se dégage ici une première piste : le descripteur, qui doit pouvoir fonctionner en contexte comme un groupe nominal, doit être hors contexte une unité lexicale susceptible d'apparaître dans un groupe nominal. Le modèle d'utilisation de la langue en indexation, ici à ses débuts, sera plus amplement spécifié dans le seconde partie de cette recherche.

(ii) Approche de la finalité de l'indexation

L'indexation nous est apparue comme une pratique destinée moins à *transmettre* de l'information, suivant une chaîne linéaire qui irait des auteurs aux utilisateurs, qu'à permettre de *construire* cette information. L'indexation ne se donne plus comme une opération qui doit déterminer, à partir de l'analyse d'un document, l'information dont l'utilisateur pourra avoir besoin. Elle agit sur un autre plan. Rapprochée de la vulgarisation scientifique, l'indexation se conçoit comme une pratique mettant en œuvre des *stratégies d'exposition* des textes, agissant plus sur les conditions d'interprétation que sur l'interprétation elle-même. En cela, elle est directement concernée moins par les « mots » eux-mêmes que par leur « mise en contexte ».

Remarquons sur ce point que l'on peut rendre compte de la finalité de l'indexation sans la concevoir en termes de recherche documentaire proprement dite. L'horizon de l'utilisateur reste présent dans notre approche mais l'utilisateur est perçu plus sous l'angle de son « rôle » (rôle d'interprète) que sous l'angle de sa « nature » (un individu singulier en quête d'information particulière).

---

[1] Voir le chapitre IV.

À plusieurs reprises, dans ce chapitre, a été évoqué l'« autour » des mots ou plus précisément ce à quoi ils réfèrent : des discours, des objets, des « choses », des contextes, etc. La question du lexique en indexation appelle en effet nécessairement celle de la référence. Nous nous proposons de préciser cet aspect de la référence en indexation dans le chapitre suivant.

CHAPITRE III
LA QUESTION DE LA RÉFÉRENCE EN INDEXATION

Comme précédemment la dimension lexicale – et sans doute plus encore –, la question de la référence[1] en indexation n'est pas, en tant que telle, présente dans le discours classique. On retrouve sur ce point une impossibilité directement liée à l'adoption implicite du modèle objectiviste (*supra*). Dans ce cadre, l'indexation ne peut en effet penser son « extérieur », ou plutôt elle ne peut le penser que sous une seule et même forme : celle du concept.

La vision de la langue et du lexique qui sous-tend l'approche classique de l'indexation va en effet de pair avec une certaine vision du monde[2]. Il importe donc, à nouveau, de dégager l'arrière-plan théorique implicite afin de faire émerger, par déplacement de modèles, ce que le discours classique ne montre pas : les processus de référenciation[3] à l'œuvre en indexation.

En dépit de son manque de visibilité, du moins dans le discours normatif[4], le rôle de la référence est pourtant déterminant en indexation. C'est Michel Le Guern et les membres de l'équipe SYDO qui ont, les premiers, attiré l'attention sur cet aspect de l'indexation. En effet, ils ont mis en valeur que, si l'indexation manipule des *mots*,

---

[1] La notion de référence peut être entendue, en première approximation, comme « la propriété d'un signe linguistique lui permettant de renvoyer à un objet du monde extra-linguistique, réel ou imaginaire ». Définition reprise du *Dictionnaire de linguistique et des sciences du langage* 1994, p. 404.

[2] Voir Rastier 1994, p. 29 : « Le maintien de la triade sémiotique garde la sémantique sous la dépendance d'une ontologie, seule capable de relier les mots et les choses, par la médiation de concepts [position mentaliste] ».

[3] *Ibid.*, p. 19 : « Ce que nous appelons ici *référence* n'est pas un rapport de représentation à des choses ou à des états de choses, mais un rapport entre le texte et la part non linguistique de la pratique où il est produit et interprété. [...] La référence ainsi définie ne relève pas de la représentation mais de l'action, telle qu'elle est structurée par une pratique ».

[4] « Le thésaurus est considéré par les spécialistes de la documentation comme une représentation symbolique ne prenant pas en compte le référent. Mais le descripteur renvoie à des documents qui traitent de la réalité, cela revient à dire que le descripteur renvoie à cette même réalité par référent interposé, c'est-à-dire par le document », Mustafa-Elhadi 1992, p. 468.

c'est pour renseigner sur les *choses* : « Les systèmes documentaires, à de rares exceptions près, ont pour finalité de fournir à l'utilisateur des renseignements sur les choses, et non sur les mots.[1] »

La problématique documentaire devient, dans cette perspective, celle de la mise en relation entre un objet du monde et le document (ou les documents) qui apportera (apporteront) des informations sur cet objet. C'est dans ce cadre que le descripteur peut être envisagé comme une unité de discours[2] : nous y reviendrons.

Nous nous inscrivons dans le cadre de l'analyse proposée par Michel Le Guern, en privilégiant un point de vue linguistique sur la référence[3]. La problématique s'exprime alors sous la forme suivante : « comment le langage parvient-il à parler du réel ?[4] », question qui aussitôt se dédouble[5] :

(i) Comment se réalise l'acte référentiel ?

(ii) Quelles sont les propriétés référentielles qui permettent aux expressions de référer ?

Cette approche de la référence ne permet certes pas de tout dire sur la question de la référence en indexation ; notamment la linguistique ne peut rien dire de la « réalité » des éléments qu'elle manipule, réalité mentale (notion de concept) ou réalité phénoménale (notion de chose). Elle permet en revanche de pouvoir concevoir l'indexation comme une opération qui crée, *via* les mots, ses propres « choses », ses propres objets : document et descripteur.

Pour montrer comment l'indexation peut se concevoir sous l'angle de la référenciation, nous procéderons de la façon suivante, en trois étapes :

- nous dégageons, dans un premier temps, la zone de tension entre modèle documentaire et modèle linguistique sur le point qui concerne l'appréhension de la référence : la vision du monde que sous-tendent les deux modèles est radicalement différente (I) ;

- nous présentons ensuite l'approche de la référence dans le cadre du modèle linguistique : la façon dont les questions sont posées ainsi que les distinctions qui y sont établies nous fournissent des repères pour identifier les phénomènes référentiels en indexation (II) ;

- nous proposons ensuite une approche de la référence en indexation qui tient compte des hypothèses linguistiques précédemment dégagées : l'indexation apparaît, dans ce cadre, comme une opération de référenciation, c'est-à-dire

---

[1] Le Guern 1991a, p. 22.
[2] « La finalité du descripteur exclut que l'on puisse l'envisager en faisant abstraction de la valeur référentielle de ses occurrences dans le corpus », Le Guern 1984, p. 164.
[3] Cependant, la référence constituant typiquement un concept scientifique interdisciplinaire, nous serons amenée à aborder quelques aspects des problématiques logique et philosophique de la référence.
[4] Kleiber 1981, p. 11.
[5] *Ibid.*, p. 13 ; les deux questions (i) et (ii) sont intimement liées : « l'acte de référence ne saurait être accompli si les expressions n'avaient pas des caractères référentiels propres ».

comme une opération qui construit ses propres objets comme ses propres effets référentiels (III).

I - Conflit entre modèles de la référence

Modèle documentaire et modèle linguistique illustrent chacun l'une des deux branches de l'alternative philosophique sur la référence.

Exprimé en termes philosophiques, le débat sur la référence fait émerger deux positions distinctes, celle des réalistes et celle des nominalistes, qui constituent la base du conflit entre les deux modèles documentaire et linguistique en présence. Après avoir rappelé les grandes lignes de ce débat philosophique (I.1), nous présenterons, d'une part, les principales marques du modèle réaliste en indexation (I.2) et, d'autre part, les limites les plus remarquables de ce présupposé réaliste pour la description des faits d'indexation (I.3).

I.1 - Les termes du débat sur la référence

Traditionnellement, l'histoire de la philosophie situe la primeur du débat sur la référence au X^{e} siècle, en pleine « querelle des universaux » qui cristallise alors l'ensemble de la réflexion philosophique sur le langage, liant à la fois logique et théologie.

Le débat philosophique sur la référence s'organise[1] autour de deux positions adverses, celle des Réalistes et celle des Nominalistes :

- pour les Réalistes, « les mots font directement référence à la réalité objective, soit permanente, soit impermanente, mais toujours réelle et réellement différenciée[2] ». C'est une position qui reprend la thèse d'Aristote[3], dite essentialiste, dans le sens où l'existence est appréhendée en termes d'essence ;

- pour les Nominalistes, la réalité n'est perceptible que par les mots : le concept n'est rien de plus qu'un nom, « qu'une construction mentale sans réalité extérieure[4] ». C'est pourquoi nous ne pouvons pas connaître la réalité en soi, dans son essence, mais seulement dans « les représentations (les dénominations) par lesquelles nous percevons les phénomènes dont nous expérimentons les sensations[5] ». C'est une position dite existentialiste, illustrée notamment par Guillaume d'Ockham, qui tient que s'il y a des propriétés communes à plusieurs éléments, on ne peut, pour autant, en inférer la réalité de ces éléments.

---

[1] Il s'agit là d'une présentation où l'on radicalise les positions de chacun ; une approche plus complète peut se trouver, par exemple, dans De Libera [1993].
[2] Zimmermann 1989, p. 402.
[3] Aristote n'illustre qu'un aspect du réalisme antique ; c'est surtout la lecture d'Aristote par les médiévaux que l'on retient ici.
[4] Zimmermann 1989, p. 402.
[5] Le Moigne 1995, p. 45.

D'option philosophique, la position réaliste s'est peu à peu imposée comme épistémologie « officielle », pour reprendre les termes de Le Moigne[1], notamment sous la forme du modèle « cartésiano-positiviste ».

C'est ainsi que la position réaliste s'est maintenue comme le modèle du sens commun, comme une *doxa* utile et utilisée dans la mesure où elle fournit un mode d'appréhension aisé et immédiat du rapport entre les mots et les choses[2]. Rien d'étonnant, là encore, à ce que le modèle documentaire épouse la forme de cette *doxa*, comme nous le verrons ci-après.

Face à ce modèle réaliste, dominant, le modèle nominaliste, moins représenté : ce sont surtout les linguistiques et la sémiologie qui, reprenant la position nominaliste, ont reconstruit une « opposition » qui prend alors le nom de constructivisme[3]. Cette épistémologie constructiviste se fonde sur les deux hypothèses suivantes : « Une hypothèse relative au statut de la réalité connaissable, qui pour être connue doit pouvoir être cognitivement construite ou reconstruite intentionnellement par un observateur-modélisateur ; et une hypothèse relative à la méthode d'élaboration ou de construction de cette connaissance qui ne fera plus appel à une "norme du vrai" (par déduction programmable) mais à une « norme de faisabilité » (par intuition re-programmable).[4] »

Cependant, si ce sont les linguistes qui ont, en grande partie, contribué à raviver le débat sur la référence[5], il n'y a pas, pour autant, consensus absolu entre les écoles : des différences se notent tout autant entre linguistes qu'entre linguistes et logiciens ou linguistes et philosophes[6]. Néanmoins, on peut formuler de façon générale la problématique de la référence en linguistique sous la forme suivante : s'il n'y a pas, dans la langue, de marque de référence, tout mot peut-il pour autant pointer indistinctement sur n'importe quel objet ? C'est principalement autour de cette question que les linguistiques abordent la question de la référence.

Sur la base de ces présupposés philosophiques, on peut faire apparaître ce qui oppose le modèle documentaire et le modèle linguistique : d'un côté, un monde construit dont il s'agit de rendre compte ; de l'autre, un monde à construire selon des procédés dont il faut rendre compte. L'ambition est radicalement différente.

---

[1] Le Moigne [*Ibid.*, p. 4-11] parle plus volontiers d'épistémologie « institutionnelle ».
[2] Voir Dubois et Mondada 1995, p. 274 : « La croyance en un monde extérieur est une propriété centrale de la "raison mondaine" ("mundane reason", Pollner 1987), qui donne une intelligibilité et une descriptivité à la réalité quotidienne, à ses représentations ordinaires, aux raisonnements de tous les jours ; qui permet en outre de traiter les contradictions ou les conflits entre des versions multiples et discordantes des "mêmes" réalités comme étant imputables à l'erreur ou à la folie ».
[3] Le Moigne 1995, p. 44-46.
[4] *Ibid.*, p. 41.
[5] Par exemple, l'hypothèse dite Sapir-Worf (chaque langue est spécifique et configure le monde à sa façon).
[6] Une présentation exhaustive du débat dont la référence a fait l'objet aussi bien en philosophie qu'en logique dépasserait largement le cadre de ce travail : on peut trouver une synthèse et des références bibliographiques dans Nef 1991, p. 85-109.

I.2 - Traces du modèle réaliste en indexation

On trouve, dans le discours classique, deux types de « trace » du modèle réaliste : l'un concerne la stabilité de la relation entre les mots et les choses (A), l'autre la préexistence des documents en cause dans cette relation (B).

A - Stabilité de la relation entre les mots et les choses

Dans le modèle réaliste, le langage est définitivement coupé du monde, comme le monde du langage : d'un côté, existe un monde discrétisé en objets, incarnations d'essences (ou concepts), de l'autre existent des mots, expressions d'essences (ou concepts). La langue est là, qui se pose comme instrument de mise en relation entre mots et choses opérant *via* les concepts.

Le monde existant préalablement à toute appréhension, il est le même pour tous les sujets parlants : c'est pour cela que l'on peut transmettre ses éléments. Le seul problème vient des mots, qui ne sont pas toujours aptes à les transmettre correctement.

Le modèle réaliste, qui suppose à la fois une stabilité des « choses » et une totale extériorité du monde par rapport au langage, constitue le modèle de « référence » de l'indexation :

- l'indexation se donne en effet comme une procédure chargée de maintenir la stabilité des choses du monde au cours de leur transfert des auteurs vers les lecteurs ;

- elle se pose à ce titre comme une simple manipulation d'objets lui préexistant – objets du monde et objets textuels[1] – qui ne sont pas, dans cette approche, distingués[2].

Dans le cadre réaliste, l'enjeu de la référence en indexation se donne essentiellement sous la forme suivante : comment assurer une stabilité du référent d'un bout à l'autre de la chaîne documentaire ? Cette question, qui se pose uniquement dans un cadre réaliste, reçoit naturellement une réponse de type réaliste. À la problématique de la stabilité référentielle correspond en effet celle de la stabilité linguistique : comme le monde est stable, il suffit de désigner ses objets toujours par les mêmes mots. C'est là le fondement même des langages documentaires : pour être sûr de désigner toujours le même objet, on doit se mettre d'accord pour employer toujours le même mot[3].

---

[1] Ou plus généralement objets documentaires.

[2] Cf. Maniez 1993, p. 254 : l'indexation se justifie parce qu'existent d'un côté des documents et de l'autre des besoins d'information. Dans ce cadre, l'indexation est ce qui doit indiquer « à un demandeur les documents relatifs *au sujet qui l'intéresse* c'est-à-dire *à la portion de la réalité sur laquelle se focalise sa curiosité* ». C'est nous qui soulignons la formulation de cette assimilation entre objet textuel et objet du monde.

[3] Cette analyse du langage documentaire comme présupposé réaliste se retrouve chez d'autres auteurs, notamment chez Turner [1990, p. 2] : « L'idée que la construction d'un langage artificiel peut conduire à réduire l'ambiguïté du langage naturel est fondée sur une *théorie positiviste du langage*. Selon ce point de vue, la signification des mots vient de ce qu'ils nomment. [...] Le travail de création d'un langage artificiel consiste à ramener la

B - Préexistence des objets documentaires

Le modèle réaliste ne permet pas de poser véritablement la question des types de référents, d'objets (mondains et/ou textuels), sur lesquels pointe l'indexation.

La question du type d'objet à indexer est en effet peu traitée dans le discours classique, comme si la problématique du document, comme celle de l'information, ne concernaient pas l'indexation elle-même.

Le document est en effet un « donné », l'« input » de l'indexation : il est en cela constitué en dehors d'elle[1]. De même l'information[2] est conçue comme étant extérieure au processus même de l'indexation ; on ne sait pas bien si elle constitue son « input » ou son « output », sans doute les deux.

On peut remarquer que le modèle réaliste imprime profondément sa marque en indexation :

- *c'est lui qui détermine les outils de l'indexation, et notamment la forme des langages documentaires : hors de ce modèle, est-on encore obligé de soutenir une corrélation entre stabilité référentielle et stabilité linguistique ?*

- *c'est lui qui autorise une mise à l'écart des objets de l'indexation du champ même de l'indexation : hors de ce modèle, il devient nécessaire de définir d'une part et de distinguer d'autre part les principaux objets de l'indexation, notamment le document et l'information.*

On voit ce que suggère une remise en cause du modèle réaliste sous-jacent au discours classique : un changement radical dans l'approche de la notion d'indexation et un important travail de définition de ses objets. Malgré le caractère ambitieux de cette tâche, il nous semble important de l'amorcer, le modèle réaliste se révélant fortement inadéquat pour rendre compte des faits d'indexation.

I.3 - Limites du modèle réaliste en indexation

Le présupposé réaliste de la préexistence d'un monde stable et déjà discrétisé, tel qu'il est assimilé dans le modèle classique de l'indexation, laisse sans réponse deux ordres de problème auxquels se trouvent régulièrement confrontées les pratiques d'indexation :

- la variabilité des objets d'indexation : l'objet d'indexation, ou encore le document, est assimilé, dans le modèle réaliste, à un objet du monde comme les autres : il existe et est prêt à être indexé. Or, pour un monde donné, est-ce que ce sont toujours les mêmes objets qui sont retenus pour être indexés ?

---

diversité de signes à un symbolisme unique. Le langage documentaire est, en théorie, une liste de symboles ayant chacun un rapport intrinsèque avec la chose qu'il représente ».

[1] La norme [Z 47-102 (1978), p. 231] appréhende en effet le document comme un objet déjà constitué : c'est l'« ensemble d'un support d'information, des données enregistrées sur ce support et de leur signification, servant à la consultation, l'étude, la preuve, etc. ».

[2] Dans les termes de la norme, l'information est entendue comme la « signification que l'homme attache aux données au moyen de conventions connues utilisées dans sa représentation » ; mais la notion de donnée y est parallèlement définie de façon circulaire comme la « représentation codée d'une information ». *Vocabulaire de l'indexation* 1987.

- la variabilité des termes d'indexation : bien que le monde soit discrétisé en entités stables et que les langages documentaires donnent de quoi représenter de façon univoque chacune de ces entités, la pratique de l'indexation se heurte encore à la variabilité des termes d'indexation[1].

Ce sont là deux aspects de la pratique d'indexation qui contredisent les présupposés réalistes adoptés dans le discours classique et qui indiquent les limites de ce modèle pour une approche des faits d'indexation. Nous précisons ces deux aspects ci-dessous.

I.3.1 - LA VARIABILITÉ DES OBJETS D'INDEXATION : LA QUESTION DU DOCUMENT

Contrairement à la définition que propose la norme AFNOR[2] et suivant plutôt une proposition de Gardin[3], nous entendrons, dans un premier temps, par document tout objet retenu pour être indexé.

Cette définition posée, la question qui se pose est la suivante : qu'est-ce qu'un objet « indexable » ? Autrement dit, comme le souligne Odile Le Guern, avant la question du choix des termes d'indexation, l'indexeur se pose bien celle du « choix des objets représentés », de « ce qu'il va choisir de montrer (index)[4] ». Dans la littérature consultée, on trouve peu de précisions sur ce point. Quand la question est évoquée, c'est pour dire qu'elle n'a justement pas été traitée[5].

Il nous semble que la question du choix des objets documentaires se pose à deux niveaux différents :

(i) qu'est-ce qui, parmi l'ensemble de la production éditoriale, préside au choix d'une source documentaire* ?

(ii) qu'est-ce qui, pour une source documentaire donnée, détermine le choix de la partie de la source effectivement indexée ?

La question (i), qui relève de ce qu'il est convenu d'appeler une « politique d'acquisition[6] », engage l'organisation de l'espace des documents dans son entier : elle sera traitée sous cet angle dans le chapitre IV.

La question (ii) renvoie à ce que l'on appelle généralement la sélection documentaire[1] : c'est elle qui constitue le document proprement dit ; c'est d'elle dont il sera prioritairement question dans ce paragraphe.

---

[1] Cf. norme Z 47-102 (1978), p. 6 : « Dans le cas idéal, pour un système donné, l'indexation d'un document devrait être identique quel que soit l'indexeur. Elle ne devrait pas non plus varier dans le temps pour un même indexeur si l'outil documentaire n'a pas été modifié ».
[2] Voir le glossaire à l'entrée « document ».
[3] Gardin [1967] propose de considérer le document comme « tout objet au sens large (objet concret, image, texte, etc.) considéré comme unité d'analyse et/ou de référence, dans les travaux d'indexation ».
[4] Odile Le Guern 1989, p. 427
[5] Suzanne Bertrand-Gastaldi [1993, p. 161] rappelle la problématique, « restée jusqu'alors sans réponse », que formule Coates en 1979 : « How does an indexer determine what is or what is not indexable ? ».
[6] Voir par exemple Calenge 1994.

Les principes de sélection les plus couramment exprimés reposent sur l'idée qu'il importe de respecter la source éditoriale telle qu'elle se donne : l'indexeur doit rester « collé » à la réalité éditoriale. Or celle-ci est, heureusement, plus malléable qu'il n'y paraît.

Pour mesurer la latitude que les indexeurs prennent par rapport à la « réalité éditoriale », nous avons réalisé une enquête auprès de dix organismes documentaires[2] pour observer comment, à partir d'un même « monde », d'une même « réalité éditoriale », différents indexeurs opéraient leur découpage en documents.

A - Expérimentation

Dans l'expérience que nous avons menée, nous entendons :
- par « source », tout article du journal *Le Monde* : toute unité textuelle typographiquement circonscrite le plus souvent entre un titre et une signature ;
- par « document », le segment textuel affecté d'un (ou de plusieurs) descripteur(s).

En examinant le rapport entre sources sélectionnées et documents indexés[3], on repère les trois cas de figure suivants :

(i) six des dix organismes documentaires sondés ont sélectionné plus de textes qu'ils n'ont indexé de documents : autrement dit, la création du document consiste ici essentiellement en un *regroupement* de différentes sources ;

(ii) deux des organismes documentaires ont indexé plus de documents qu'ils n'ont sélectionné de sources : dans ce cas, la même source subit un *éclatement* en objets documentaires différents ;

(iii) seuls deux des dix organismes documentaires étudiés ont indexé autant de documents qu'ils avaient sélectionné de sources : autrement dit, le type de source sélectionnée constitue, pour ces organismes, un objet d'indexation pertinent.

Cette première observation montre que l'indexation ne manipule pas toujours les objets textuels (les sources) tels qu'ils lui préexistent mais qu'elle crée ses propres objets.

Ce constat est renforcé par un examen plus minutieux qui montre que, à partir d'une même source (un même article *du Monde*), les différents organismes documentaires créent des documents différents.

La diversité des cas de figure est bien représentée par le traitement documentaire de deux articles du *Monde* portant, pour le premier, sur la chaîne de télévision Arte et,

---

[1] Ou encore la « couverture documentaire » : ces notions renvoient aux « domaines » que traitent prioritairement une bibliothèque ou un centre de documentation. Cet aspect de la sélection documentaire reste le plus souvent à la discrétion des organismes documentaires : on ne trouve que peu de règles formalisées expliquant les choix effectués. Nous y revenons dans le chapitre IV.

[2] L'annexe 1 détaille les participants et les consignes de l'expérience réalisée.

[3] Voir annexe 2.

pour le second (un encadré) sur Jean-Marie Cavada (alors président de La Cinquième, promu président du GIE Arte-La Cinquième)[1] :

- soit les deux articles constituent chacun un document ;

- soit un seul des deux articles est retenu comme document ;

- soit les deux articles sont regroupés en un document ;

- soit les deux articles sont regroupés dans une même revue de presse (comportant d'autres « sources ») ; et c'est elle qui constitue un document (qui est affectée de descripteurs) ;

- soit l'un des deux articles est intégré dans une revue de presse qui, elle, forme un document particulier, indexé en tant que tel.
  On observe également une variabilité de « mise en document » pour une même source au sein d'un même organisme documentaire ; ainsi le centre de documentation de la *Fondation nationale des sciences politiques* procède-t-il en deux temps :

  - dans un premier temps, une source du *Monde* est constituée en document, c'est-à-dire, pour cet organisme documentaire, sélectionnée, affectée d'un indice de classification et intégrée dans un dossier documentaire ;

  - dans un second temps, c'est le dossier documentaire lui-même, composé de coupures de presse de source hétérogène, qui constitue à son tour un « document » : il est, dans son intégralité, indexé dans un autre langage documentaire (Rameau en l'occurrence) et intégré comme *une* seule « référence » (au sens bibliothéconomique du terme) dans le catalogue de la bibliothèque.

Ces différentes observations, issues d'une expérience dont la portée exemplaire n'est, par ailleurs, pas prouvée[2], montrent cependant que le document, loin d'être un « donné », est un « construit », ou, mieux, un « état ». Il semble, en effet, que tout texte ou ensemble de textes peut fonctionner, à un moment donné, comme « un document[3] ».

L'expérience menée nous fournit une parfaite illustration de ce dernier point (le document conçu par « intention »). C'est à nouveau la pratique d'indexation du centre de documentation de la *Fondation nationale des sciences politiques* qui illustre ce cas de figure. A été retenu, en effet, comme unité d'indexation, un « non-article », une « lettre ouverte à monsieur Jacques Chirac » signée d'une association de « professeurs et accompagnateurs des conservatoires municipaux de Paris[4] », qui se rapproche d'une publicité (au sens premier du terme) et qui n'est assurément pas

---

[1] *Le Monde* 1/12/1994, p. 10.
[2] Il faudrait qu'une expérimentation similaire soit effectuée à partir d'autres types de sources, issues par exemple de la littérature scientifique et technique. Retrouve-t-on, sur des domaines « étroits », cette même variabilité dans la « mise en documents » à partir d'une même « source » ?
[3] Voir, sur ce point, la notion de « document par intention » [Benoît (dir.) 1992] : « est document tout support d'information voulu comme tel ».
[4] *Le Monde* 1/12/1994, p. 14.

un article. Ce dernier exemple souligne ce qui ressortait déjà des observations précédentes : le document se présente comme un texte (ou ensemble de textes) en usage, ou plutôt mis en usage.

Les données de cette expérimentation permettent de montrer, au moins sur le cas du traitement documentaire de la presse généraliste, que, pour une « réalité éditoriale donnée », il y a toujours plusieurs mises en document possibles et qu'à ce titre, le document ne peut être considéré comme un donné : il est nécessairement construit et construit par « intention ».

B - Conclusions de l'expérimentation et formulation d'hypothèses

L'ensemble des précédentes observations issues de notre expérimentation amène à formuler un ensemble de questions et d'hypothèses relatif au document d'une part et à l'indexation d'autre part.

Questions et hypothèses relatives à l'approche du document en indexation

Nous ferons l'hypothèse que le document est en indexation un « construit », ou plutôt un « reconstruit » à partir d'une source, dont il propose une utilisation : nous nous placerons dans une approche constructiviste du monde.

- Les questions à résoudre seront alors de divers ordres :

- en quoi consiste la construction du document ? Qu'est-ce que le document retient de la source et qu'est-ce qu'il ne retient pas ? La diversité de la « mise en document » des sources induit-elle que l'« information » proposée est différente d'un organisme documentaire à l'autre ? Autrement dit l'opération de mise en document intervient-elle dans l'interprétation de la source proposée à l'utilisateur ?

- parmi l'ensemble des sources existantes, qu'est-ce qui guide le choix d'une source ? Sur quels critères une source peut-elle être utilisée comme un document ? Quelles sont les propriétés que doit avoir une source pour fonctionner comme document ?

- quel est le rapport entre le monde des choses et le monde des documents créés en indexation ? En quoi les documents, plus que les « sources », sont-ils à même de renseigner un utilisateur sur les « choses », pour reprendre les termes de Michel Le Guern ?

- Questions et hypothèses relatives au processus d'indexation

Si l'indexation ne porte pas sur des objets existants mais qu'elle construit ses propres objets, il semble important de la redéfinir en intégrant cette dimension de la création du document : on dira à ce titre que le processus de l'indexation comprend une phase de mise en document. Cette phase, classiquement abordée sous l'angle de la sélection documentaire, est traditionnellement conçue comme distincte de celle de l'indexation.

Cependant, la proximité entre ces deux opérations a pu être relevée, par Le Loarer par exemple : « Le processus d'indexation peut être lié à (ou suivre) celui qui est relatif à la sélection des documents à indexer : parmi les ouvrages reçus ou acquis dans un centre de documentation ou une bibliothèque, quels sont ceux qui font l'objet d'une indexation ? dans un fascicule de périodique, indexe-t-on tous les articles ou seulement certains ? dans des actes de congrès traite-t-on toutes les communications ou seulement certaines ?[1] »

Nous faisons ici hypothèse que l'opération d'indexation inclut une phase de sélection et de construction de ses objets.

Cette phase est réalisée *a priori* sur la base d'une présomption, qui relève moins d'un souci des besoins de l'utilisateur que d'une « vision du monde » que l'on estime pertinente à fournir à un utilisateur[2]. S'il est vrai que l'indexation doit se penser dans les termes d'une « analyse prévisionnelle[3] », c'est moins *en aval* de l'indexation, au niveau des termes à choisir, qu'*en amont* de l'indexation au niveau des sources à retenir et des documents à construire.

Notre approche de l'indexation nous conduit à reformuler la notion de prédiction telle qu'elle se conçoit habituellement en indexation[4] : elle n'affecte plus les « mots[5] » mais les « choses ». L'indexation apparaît en ce sens comme un acte de discrétisation du monde qu'elle propose de faire voir sous un certain angle (celui des documents qu'elle retient et qu'elle organise). Dans ce cadre, et en opposition au modèle réaliste, la variabilité des objets d'indexation n'est plus un problème, mais au contraire la marque distinctive de l'indexation : on attend de l'indexation qu'elle puisse donner un sens à la partie du « réel » qu'elle choisit de montrer[6].

1.3.2 - VARIABILITÉ DES TERMES D'INDEXATION : LA QUESTION DE LA STABILITÉ RÉFÉRENTIELLE

La variabilité des termes d'indexation[7], c'est-à-dire, par exemple, le constat que deux indexeurs d'un organisme documentaire donné choisissent, parmi les termes d'un même langage documentaire, deux descripteurs différents pour indexer un même document, ne pose un problème que dans le cadre réaliste. Or c'est précisément dans ce cadre qu'il ne peut être résolu.

---

[1] Le Loarer 1994, p. 152.
[2] C'est là où l'indexation rejoint la problématique de la catégorisation qui établit une discrétisation en « domaines » ; la détermination de ces domaines repose nécessairement sur une certaine perception d'un monde, sur une « idéologie », voir sur ce point Escarpit 1991, p. 156.
[3] *Ibid.*, p. 167.
[4] La notion de prédiction est surtout pensée dans le discours classique en termes de « représentation » lexicale du « contenu ». Ainsi Fugmann [1993], qui a particulièrement développé cet aspect au point d'en faire un axiome de l'indexation, parle-t-il de la « prédictibilité de la représentation ».
[5] Toute la problématique des langages contrôlés relève de la croyance en la nécessité et en la possibilité de *prévoir la forme linguistique* la plus susceptible d'être utilisée à la recherche.
[6] C'est la notion d'« intention de sens » que défend Batime pour la construction des systèmes d'information. Batime 1995, p. 19-25.
[7] Si le problème de la variabilité en indexation est beaucoup plus large [voir Bertrand 1993, pour une revue de la question], la variabilité dans les termes d'indexation constitue néanmoins l'aspect le plus étudié (selon Bertrand 1993, p. 20).

En effet, la notion de variabilité de l'indexation repose sur une mesure de la disjonction entre stabilité référentielle (incarnée dans le document, vu comme donnée stable pourvue d'une signification[1]) et stabilité linguistique (incarnée dans le descripteur, représentant lui aussi d'un sens unique[2]). Que stabilité référentielle (stabilité des « choses ») et stabilité linguistique (stabilité des « mots ») se correspondent relève d'une conception réaliste de la signification[3] qui ne peut expliquer les cas de disjonction qu'en termes de faute ou d'erreur[4].

Hors d'un cadre réaliste, la problématique de la stabilité référentielle ne se pose pas, ou du moins ne se pose plus dans les mêmes termes[5]. Elle apparaît non plus sous l'angle de la production (il ne s'agit plus de produire une unité linguistique « stable » pour exprimer toujours le « même » référent) mais sous l'angle de la réception ; d'un point de vue interprétatif peuvent se créer des effets de stabilité référentielle[6]. Là encore, dans le cadre réaliste, où le monde est donné comme stable et discrétisé en objets de la même façon pour tous les sujets parlants, la différence des points de vue production/indexeur et réception/utilisateur ne peut apparaître. Or ce point de vue est déterminant quand il s'agit de penser la notion de stabilité référentielle en indexation : il permet notamment de distinguer fin et moyens.

Si l'indexation peut se donner pour fin d'assurer une relation référentielle stable, cet objectif se réalise nécessairement au niveau de la réception (de l'utilisateur) et non à celui de la production (de l'indexeur). En effet, l'indexeur, qui ne dispose que d'unités lexicales hors emploi, ne peut à proprement parler établir de relation référentielle, celle-ci ne se créant qu'en discours : « Les noms, en effet, ne trouvent pas leurs référents tout prêts dans la nature, mais ils doivent pour ainsi dire les construire à chaque fois dans la communication, en découpant les classes d'objets dans le tissu de l'expérience.[7] »

La question de savoir si c'est toujours le même référent que le mot construit « dans la communication » ou bien si celui-ci varie en fonction de chaque discours, de chaque contexte pourra être abordée dans le cadre d'une approche linguistique de la référence. En effet, de ce point de vue, on peut interroger la morphologie du descripteur.

Les noms propres sont, comme nous le verrons, particulièrement à même d'établir une relation référentielle stable. Sur ce point, il n'est pas indifférent que les

---

[1] Voir le glossaire pour la définition normative du document.
[2] *Id.*
[3] Apothéloz et Reichler-Béguelin 1995, p. 202 : une conception réaliste de la signification est une conception « dans laquelle le signifié se réduirait à une relation rigide non manipulable par les sujets parlants, entre la langue et le monde ».
[4] Voir par exemple la remarque de Dubois 1995, p. 90 : les erreurs, décalages, difficultés ou inadéquations rencontrés par les professionnels sont imputés à la fois aux outils (rustiques mais perfectibles) et aux humains « à l'évidence condamnés aux imperfections ».
[5] Dubois et Mondada 1995, p. 282 : « Notre argument consiste à dire que la "stabilité" résulte en fait d'un point de vue réaliste qui relie les catégories à des propriétés du monde – comme si l'objectivité du monde produisait la stabilité des catégories – au lieu de les relier à des discours socio-historiques et à des procédures ancrées culturellement ».
[6] Voir Dubois et Mondada [*Ibid.*, p. 273 notamment] qui parlent d'« effet stabilisateur des pratiques ».
[7] Formigari 1992, p. 448.

premières pratiques d'indexation se soient constituées autour du nom propre[1] ; mais il n'est pas indifférent, non plus, que les pratiques d'indexation aient plus tard utilisé aussi des noms communs, qui, eux, ne peuvent, en règle générale, établir de relation référentielle stable.

C'est donc du point de vue de la réception, de l'interprétation que se créent, notamment par le biais des noms propres des *effets* de stabilité référentielle : « S'agissant de "personnages", ou tout autre objet-de-discours susceptible d'être redésigné sur la durée par le même nom propre ou par un désignateur peu contingent, on peut supposer que les opérations de référence sont au bénéfice d'un statut cognitif spécial. Du fait même des propriétés sémantiques des désignateurs dits rigides, un tel objet-de-discours *apparaîtra comme stable* en tant qu'objet mémoriel : son identité pourra *donner l'illusion* d'être indépendante des prédications dont il fait l'objet, quand bien même les connaissances qu'on a de lui évoluent au fil du discours.[2] »

La variabilité des termes d'indexation apparaît dans le discours classique comme un problème sans réelle solution : on finit par adopter une position fataliste où l'on invoque immanquablement l'imperfection humaine. Or ce problème ne se pose que dans le cadre du modèle réaliste implicitement adopté.

I.4 - Conclusions intermédiaires

Au début de ce paragraphe, nous cherchions à dégager la « zone de tension » entre modèle documentaire et modèle linguistique sur la question de la référence. En reprenant, de façon schématique, l'alternative classique en philosophie entre Réalistes et Nominalistes, nous avons proposé d'identifier la zone de conflit entre les deux modèles en termes de vision du monde : dans le modèle documentaire classique d'inspiration réaliste, le monde est donné ; dans le modèle linguistique d'inspiration nominaliste, le monde est construit.

Au-delà de simples positions de principe, quelle est l'incidence de l'adoption de l'un ou l'autre des deux modèles ?

Nous avons cherché à montrer que l'adoption implicite du modèle réaliste en indexation se révélait, pour qui cherche à analyser les faits d'indexation, à la fois limité et opacifiant : limité, dans le sens où ses présupposés se trouvent pris en défaut par la pratique sans que le modèle ne puisse proposer ni une explication ni une solution ; opacifiant, dans la mesure où fin et moyens de l'indexation sont, dans le modèle réaliste, peu distingués, ce qui conduit au « bricolage » de langages documentaires conçus pour réaliser une double stabilité, référentielle et linguistique, alors même qu'il s'agit là d'effets de nature différente. L'effet de stabilité linguistique du descripteur relève du lexique (elle vient de son autonomie lexicale) ; l'effet de stabilité référentielle relève du discours (elle vient de l'interprétation).

Pour concevoir les différents aspects de la référence en indexation – ceux qui concernent le document d'une part et ceux qui concernent l'effet de stabilité

---

[1] Voir Escarpit 1991, p. 153 : les catalogues d'Alexandrie offraient des entrées qui ont été d'abord les premiers mots d'un document « puis plus tard le nom de l'auteur quand l'imprimerie a tiré l'œuvre de l'anonymat ».
[2] Apothéloz et Reichler-Béguelin 1995, p. 266 (c'est nous qui soulignons).

référentielle d'autre part –, nous devons au préalable disposer d'un modèle qui nous permette de traiter ces questions de façon explicite. C'est à ce titre que nous sollicitons une approche linguistique de la référence (paragraphe II, suivant), qui nous permettra de proposer une approche documentaire explicite de la référence (paragraphe III de ce chapitre).

II - L'approche de la référence dans le modèle linguistique

Le modèle linguistique[1] formule la problématique de la référence sous un angle qui, en distinguant les objets et en reformulant les questions, nous permet d'approcher de façon plus précise les phénomènes référentiels en indexation.

Nous présentons succinctement la formulation de la problématique de la référence en linguistique, puis nous nous focalisons ensuite sur la question centrale de la référence en linguistique : celle du rapport entre le sens et la référence.

II.1 - Formulation de la problématique de la référence en linguistique

Comment nous l'avons mentionné dans le premier chapitre et comme il est apparu dans le chapitre II[2], la méthodologie d'analyse en linguistique ne permet pas de traiter directement de faits massifs comme celui de la référence.

La question de la référence est en linguistique généralement abordée sous l'angle de phénomènes de langue précis et jugés exemplaires : c'est le plus souvent par l'étude des anaphores que sont posés les problèmes de référence. Nous esquissons rapidement ce cadre privilégié de l'étude de la référence en linguistique. Nous présentons ensuite les principales options méthodologiques de l'analyse linguistique de la référence : il s'agit, d'une part, de distinguer les objets traités (types de référents et acte de référence) et, d'autre part, de sérier les aspects de la problématique de la référence que la linguistique peut traiter (qu'est-ce que la linguistique peut dire de la référence en restant dans son domaine de compétence ?).

---

[1] L'emploi du singulier ne signifie pas qu'il y a consensus, en linguistique, sur l'approche de la référence ; sur ce point Kleiber 1981 donne, nous semble-t-il, une bonne vue d'ensemble des différentes positions linguistiques sur la référence. Reste que pour l'opposition que nous traitons dans ce chapitre, nous pouvons aborder l'approche linguistique de la référence dans son ensemble puisque, quelles qu'elles soient, les positions linguistiques contemporaines se situent toutes dans un cadre non réaliste, même si elles ne le font pas exactement dans les mêmes termes, voir notes 53 et 54. Sur ce point, Apothéloz et Reichler-Béguelin 1995, p. 240 : « Le réalisme n'intéresserait en soi que la critique philosophique si, au plan linguistique, il ne conduisait en droite ligne à des conceptions aujourd'hui intenables en matière de sémantique lexicale. [...] Cette position va forcément de pair avec une conception nomenclaturiste du lexique dont Saussure a fait la critique que l'on sait ».

[2] Dans le chapitre II (§ II.2.1), nous avons vu que, dans le cadre retenu, l'analyse de la signification lexicale se faisait sur des pans délimités du matériel lexical, non pas mot par mot, mais par type de mots ; à ce titre, nous avons pris, comme exemple d'analyse de la signification lexicale, le domaine que travaille D. Corbin : celui des mots construits.

II.1.1 - CADRE PRIVILÉGIÉ POUR L'ÉTUDE DE LA RÉFÉRENCE EN LINGUISTIQUE

Qu'il s'agisse de Milner [1976], de Corblin [1995] ou encore d'Apothéloz et Reichler-Béguelin [1995], c'est en général le problème de l'anaphore qui est retenu pour aborder la question de la référence[1]. Parmi les différentes problématiques qui relèvent d'une analyse des anaphores, c'est celle des anaphores « évolutives » qui semble le mieux poser les problèmes de la référence.

Pour simplement situer le cadre des questions que pointent les anaphores évolutives, prenons l'exemple le plus « fameux » de l'emploi des anaphores évolutives, celui de la recette du poulet : « Tuez un poulet actif et bien gras. Préparez-le pour le four. Coupez-le en quatre morceaux et faites le rôtir avec du thym pendant une heure. »

La question qui est posée à partir de ce texte est la suivante : à quoi renvoient les quatre anaphores « le » ? D'un point de vue strictement grammatical, les quatre anaphores « le » devraient renvoyer à la première mention du groupe nominal « un poulet actif et bien gras » ; du point de vue intuitif de la compréhension, les quatre anaphores « le » renvoient plutôt à des états successifs de l'objet « poulet ». La question est de savoir ce qui « évolue » au cours de ce texte : le référent-objet « poulet » ou bien le référent-mot « poulet » ?

Encore faut-il poser une distinction entre ces deux types de référent et, à partir de là, se demander ce qui ressort du travail du linguiste : l'examen de l'évolution des choses ou bien l'examen de l'évolution de la saisie des choses par les sujets parlants.

C'est en ce sens que l'étude précise des phénomènes particuliers d'anaphore peut permettre de tenir des propositions plus générales sur la référence d'un point de vue linguistique, comme le rappellent clairement Apothéloz et Reichler-Béguelin : « La problématique des référents évolutifs ne trouve à nos yeux d'intérêt qu'à condition d'être replacée au sein de celle, plus générale, de l'évolution de la référence et de la catégorisation ; il s'agit alors d'envisager la globalité des paramètres qui conditionnent la gestion de l'acte référentiel par un sujet plongé dans une situation de communication concrète.[2] »

Il va de soi que les distinctions à faire (les différents types de référent) et que les questions à sérier (relatives à l'évolution de la référence) ne peuvent s'exprimer que dans un cadre philosophique non réaliste : est-ce pour autant un cadre d'analyse résolument nominaliste ? Les propos de Milner sont sur ce point plus que nuancés[3] ; ceux de Apothéloz et Reichler-Béguelin ne sont pas dénués d'ambiguïté[4].

---

[1] Par exemple, Apothéloz et Reichler-Béguelin 1995, p. 243-244 : « Les virtualités innovantes de l'acte référentiel transparaissent bien dans l'emploi des anaphores lexicales, domaine où la latitude de choix dans les moyens linguistiques utilisés est accrue par le fait même que l'objet désigné est déjà identifié et en général dénommé dans le modèle du monde construit par le discours ».

[2] *Ibid.*, p. 266.

[3] La question est abordée entre autres dans Milner 1989, principalement aux pages 67-68, 138-139, 168-169.

[4] « Cette option théorique [non réaliste] ne signifie pas bien entendu que, pour nous, les échanges langagiers se dérouleraient uniquement au plan d'une sémiose déconnectée de la

Ne pouvant entrer dans la délicatesse de ces débats, nous considérerons que l'étude linguistique de la référence s'énonce dans un cadre philosophique non réaliste (au sens « médiéval » du terme, tel qu'abordé au § I.1 de ce chapitre).

II.1.2 - DISTINGUER LES OBJETS : LES DIMENSIONS DU RÉFÉRENT

Que faut-il entendre par « référent », par « objet extra-linguistique » ?

En première approximation, nous avions proposé de comprendre la référence comme « la propriété d'un signe linguistique lui permettant de renvoyer à un objet du monde extra-linguistique, réel ou imaginaire[1] ». Cette définition présente l'inconvénient de pousser à identifier l'opposition domaine linguistique / domaine extra-linguistique à l'opposition mot / chose ; Berrendonner appelle à plus de nuance : « On peut considérer le monde extra-linguistique non plus comme un référent absolu mais comme le lieu de la manifestation du sensible, susceptible de devenir la manifestation du sens humain, c'est-à-dire de la signification pour l'homme, et traiter en somme le référent comme un ensemble de systèmes sémiotiques plus ou moins implicites.[2] »

En effet, il est tout à fait possible de postuler la référence, c'est-à-dire postuler que la langue parle DE (quelque chose), sans doute même doit-on le faire[3], sans pour autant postuler l'existence, ou encore la réalité ontologique, de ce dont on parle.

Il faut pour cela établir les différentes dimensions du référent, selon le point de vue de l'analyse considéré. Notamment, il importe de distinguer référents mondains d'une part et référents discursifs d'autre part[4] :

- la notion de référent mondain correspond aux choses extra-linguistiques de la réalité mondaine[5] ;

- la notion de référent discursif capte la notion de « représentation » que les sujets parlants se font des choses mondaines *via* l'activité langagière. Il est d'usage, notamment pour marquer la distinction avec la notion classique de référent, de préférer alors le terme « objet-de-discours » : cette notion empruntée au logicien Grize peut être entendue comme une « représentation alimentée par l'activité langagière[6] ».

---

réalité, où la notion de référent se trouverait purement et simplement évacuée ou – ce qui revient au même – identifiée au signifié linguistique. D'une part, nous pensons que l'identité des objets-de-discours intègre forcément certains paramètres référentiels (au sens extensionnel du terme) ; d'autre part, il est bien entendu que l'interprétation des expressions référentielles sollicite constamment notre connaissance et notre expérience des propriétés du monde "réel" ». Apothéloz et Reichler-Béguelin 1995, p. 240.

[1] Définition reprise du *Dictionnaire de linguistique et des sciences du langage* 1994, p. 404.

[2] Berrendonner 1978, p. 21.

[3] Voir Bonhomme 1987, p. 31 : sans la référence, « le langage se dissout dans le non-sens car, comme le dit excellemment Benveniste, parler revient toujours à parler DE. C'est précisément dans ce DE que se déploie la référence qui procure au langage des objets (au sens logique) sur lesquels il fonctionne et qui établit par la suite de nouveaux objets de discours ».

[4] Apothéloz et Reichler-Béguelin 1995, p. 239.

[5] *Id.*

[6] *Id.*

Reste qu'existe nécessairement un lien entre un référent mondain et les représentations dont il est l'objet. Cependant ce lien n'est pas évident à capter, surtout au niveau d'où se place le linguiste. En effet, dès qu'ils entrent dans un discours, les référents (ou encore les *realia*) deviennent nécessairement des objets de discours : « C'est qu'une fois promus au statut d'objets-de-discours, ou assimilés à une quelconque pratique sociale, l'identité [des] *realia* devient le produit d'une interaction entre le sujet humain et son environnement. On ne peut plus dès lors se contenter de parler d'eux uniquement comme des *référents* au sens mondain du terme, dans la mesure où ces objets ont acquis le statut de construits culturels, et où par conséquent leur "essence" comporte forcément un paramètre anthropologique.[1] »

Ainsi l'objet d'analyse du linguiste ne peut être que l'objet de discours, sans quoi il cesse de faire de la linguistique et « court le risque de s'égarer dans une recherche sans fin sur ce que sont ou pas les essences des *realia* ». Cette recherche relèverait typiquement du champ philosophique : « Quant à savoir quels attributs sont constitutifs de l'identité profonde des réalités désignées, lesquels peuvent être modifiés ou supprimés sans que cette identité soit atteinte voire détruite, lesquels relèvent d'une identité "qualitative", "individuelle" ou "sortale". [...] Il s'agit là de problèmes philosophiques. Nous nous estimons quant à nous incompétents pour spéculer sur l'"essence" des *realia* susceptibles d'entrer à titre d'objets dans les pratiques langagières.[2] »

Dans ce cadre d'analyse de la référence ainsi comprise, les objets-de-discours, dont on fait l'hypothèse qu'ils entretiennent un lien avec les référents mondains, ne préexistent pas au discours : créés uniquement par le discours, ils n'ont à ce titre aucune stabilité référentielle (de désignation) ; l'instabilité est proprement constitutive de ce type d'objet : « Les dits objets-de-discours ne préexistent pas "naturellement" à l'activité cognitive et interactive des sujets parlants, mais doivent être conçus comme produits – fondamentalement culturels – de cette activité.[3] »

C'est ainsi que ce modèle d'analyse montre que, ce qui évolue, dans les cas d'anaphores évolutives, ce n'est pas le référent mondain, c'est l'objet-de-discours, mais c'est là sa propriété même ; autrement dit, la question des anaphores évolutives doit être reformulée : « Force est de constater que le problème des référents évolutifs n'en est pas un : tout objet-de-discours est, par définition, évolutif, car chaque prédication le concernant modifie son statut informationnel en mémoire discursive.[4] »

II.1.3 - SÉRIER LES QUESTIONS : LES PROBLÉMATIQUES LINGUISTIQUES DE LA RÉFÉRENCE

Compte tenu, d'une part, de la distinction établie entre référent mondain et référent discursif et, d'autre part, de la propriété dégagée d'instabilité des objets de discours, comment peut se formuler la problématique de la référence en linguistique ? On dégage ci-dessous quelques-unes des questions que la linguistique se propose de traiter sur la question de la référence.

---

[1] Apothéloz et Reichler-Béguelin 1995, p. 239.
[2] *Id.*
[3] *Ibid.*, p. 229.
[4] *Ibid.*, p. 240.

(i) Concernant l'étude du référent discursif, il ne peut s'agir, d'un point de vue linguistique, que d'une étude de l'évolution de la référence, c'est-à-dire des stratégies que mettent en œuvre les sujets parlants pour faire évoluer les objets de leur discours[1]. Ce type d'étude relève de ce que Rastier a proposé d'appeler la référenciation, entendue comme « le rapport entre le texte et la part non linguistique de la pratique où il est produit et interprété[2] ». Ce type d'approche s'oppose aux approches classiques de la référence, comme le signalent Dubois et Mondada : « Au lieu de partir du présupposé d'une segmentation *a priori* du discours en noms et du monde en entités objectives, et ensuite, de questionner la relation de correspondance entre l'une et l'autre – il nous semble plus productif de questionner les processus de discrétisation eux-mêmes.[3] »

L'étude du *processus de discrétisation* consiste à analyser la façon dont les sujets parlants construisent leurs objets (objets de discours sur un plan linguistique, catégories sur un plan cognitif[4]).

(ii) De façon corollaire, une analyse en termes de référenciation engage à une analyse des *processus de stabilisation* obtenus au cours de l'activité des sujets parlants. L'instabilité reste la donnée de départ à partir de laquelle sont étudiés les effets de stabilité référentielle : « Nous aimerions en outre souligner qu'au lieu de présupposer une stabilité *a priori* des entités dans le monde et dans la langue, il est possible de reconsidérer la question en partant de l'instabilité constitutive des catégories à la fois cognitives et linguistiques ainsi que leurs processus de stabilisation.[5] »

L'étude des effets de stabilisation se place, non plus du côté du sujet parlant qui produit et contrôle son discours, mais du côté de l'interlocuteur ou du lecteur qui reçoit, interprète le discours[6].

L'approche linguistique de la référence, qui situe généralement son point de départ dans l'étude des anaphores, propose une méthode d'analyse qui distingue les dimensions du référent, et notamment les dimensions mondaine et discursive, et qui circonscrit les marges du problème : du point de vue de la production, il s'agit d'étudier la construction de la référence, les modalités de discrimination du monde ; du point de vue de la réception, il s'agit d'étudier les effets de stabilisation obtenus par l'activité des sujets parlants au sein d'une pratique.

L'approche linguistique de la référence accorde donc une place centrale à l'acte de référence, acte à partir duquel sont étudiées les propriétés référentielles proprement dites des expressions linguistiques[7]*. Sur ce dernier point, la question centrale pour le linguiste consiste à déterminer sur quelle base se construit la référence, l'objet de discours. La problématique est alors la suivante : si la*

---

[1] Apothéloz et Reichler-Béguelin 1995, p. 264-265.
[2] Rastier 1994, p. 19.
[3] Dubois et Mondada 1995, p. 275.
[4] *Ibid.*, p. 276.
[5] *Ibid.*, p. 275.
[6] Apothéloz et Reichler-Béguelin 1995, p. 266-267 : « Mais il faut bien voir que l'effet de coréférence [l'un des aspects de la stabilité référentielle] résulte alors davantage des investissements interprétatifs du décodeur, que des transformations subies ou non subies, concrètement ou sémiotiquement, par le référent discursif ».
[7] Comme le souligne Kleiber [1981, p. 13 par exemple], il importe de distinguer, dans la construction de la référence, l'acte de référence lui-même et les moyens par lesquels il s'effectue.

référence se construit sur une base lexicale, comment rendre compatible la stabilité de la signification lexicale avec l'instabilité des objets de discours qu'elle permet de construire ?

II.2 - La question du rapport entre sens et référence

La question du rapport entre sens et référence concerne non pas l'acte de référence lui-même mais les expressions linguistiques susceptibles de contribuer à la réalisation de cet acte. Cette question se situe donc en amont de l'acte référentiel et par conséquent en amont de la constitution des objets de discours proprement dits.

Après avoir présenté les enjeux que sous-tend la question du rapport entre sens et référence, on propose, en reprenant les arguments de Milner, de considérer le groupe nominal comme l'atome référentiel minimal. Pour finir, on donne deux exemples d'analyse linguistique qui montrent comment la référence se construit sur la base de la signification lexicale.

II.2.1 - ENJEUX DU RAPPORT ENTRE SENS ET RÉFÉRENCE

L'enjeu du rapport entre sens et référence s'inscrit dans le cadre des problématiques suivantes :

(i) quelles sont les expressions linguistiques dotées de pouvoir référentiel ? De façon générale, les linguistiques s'accordent pour voir dans le groupe nominal l'atome référentiel minimal ;

(ii) si la construction de la référence s'effectue en discours, est-elle entièrement dépendante du discours, c'est-à-dire complètement déconnectée de la langue ? Sur ce point, il n'y a pas de consensus sur ce qui déclenche la construction référentielle.

On peut en effet dégager deux principales positions :

(i) pour les uns, ce sont les pratiques sociales qui sont à l'origine de la création de la référence. C'est une position défendue par exemple par Rastier[1]. De façon moins nette, Reichler-Béguelin et Apothéloz soutiennent eux aussi cette hypothèse[2] ;

(ii) pour les autres, comme Milner, dans la mesure où n'importe quel mot ne permet pas de construire n'importe quel objet de discours[3], on est obligé de postuler que « quelque chose » guide la construction du référent ; on fait l'hypothèse que ce « quelque chose » est la signification lexicale. C'est dans ce cadre que se pose la question du rapport entre sens et référence et l'essentiel de ses problématiques. En effet, compte tenu d'une part de la stabilité de la signification lexicale et d'autre part de l'instabilité des objets de discours, on doit supposer que le sens des unités lexicales est sous-déterminé

---

[1] Rastier [1994].
[2] Apothéloz et Reichler-Béguelin 1995, p. 266.
[3] Ou alors c'est la langue elle même qui est niée : « De même qu'un Œdipe libre d'épouser sa mère, une langue où tout pourrait se dire est une contradiction dans les termes », Milner 1978, p. 10.

> par rapport à la référence, autrement dit que la signification d'une unité lexicale n'épuise pas ses possibilités interprétatives et référentielles[1].

La question du rapport entre sens et référence devenant celle d'un rapport de sous-détermination dessine du même coup un espace de « jeu » interprétatif, espace dans lequel peuvent intervenir les pratiques sociales.

La construction de la référence établie sur la base de la signification lexicale nous paraît en cela constituer un modèle pertinent permettant de penser à la fois le rôle discriminant des expressions linguistiques dans le découpage référentiel (n'importe quelle expression linguistique ne construit pas n'importe quel référent) et le rôle discriminant des pratiques sociales dans la fixation de la référence proprement dite, les pratiques sociales intervenant mais à un autre niveau et de façon seconde.

II.2.2 - L'ATOME RÉFÉRENTIEL MINIMAL : LE GROUPE NOMINAL

De façon générale, les linguistes s'accordent à reconnaître que seuls les groupes nominaux[2] (terme abrégé par GN désormais) réfèrent. Chaque école linguistique propose une approche particulière du GN : nous en aborderons deux dans le chapitre V, lorsqu'il apparaîtra nécessaire d'examiner les propriétés que l'on peut attribuer au GN.

À ce stade de l'exposé, on ne définira pas le GN à proprement parler. On en décrira simplement l'apparence[3] en disant qu'un GN, c'est une séquence dont la forme minimale est : SPEC (N), où SPEC correspond à un spécifieur (qui peut être de type article, démonstratif ou quantifieur) et où N est une unité lexicale de catégorie nominale qui, en position de « tête » dans le groupe, lui donne son nom : groupe nominal. L'unité lexicale de catégorie N peut se voir complétée par des groupes adjectivaux, des groupes prépositionnels et des propositions (relatives et complétives). Seront donc identifiées comme des GN les séquences suivantes : « cet homme, le grand homme, un homme vert, l'homme qui rit, trois hommes aux pistolets d'or, un grand homme vert qui rit, etc. ».

Le GN est un individu linguistique qui n'existe qu'en discours (il est construit par la syntaxe) et s'oppose en cela à l'unité lexicale de catégorie N (comme « homme ») qui, elle, appartient au lexique. La question du rapport sens et référence se pose donc à travers les relations entre une unité lexicale de catégorie N et le GN dont elle constitue la « tête » : la première de ces unités est dotée d'une signification lexicale (nécessairement stable), le second constitue un objet de discours (nécessairement instable).

Pour capter le rapport qu'entretiennent N et GN du point de vue de la référence, nous retiendrons, malgré les critiques dont elle peut faire l'objet[4], la distinction qu'établit Milner entre « référence actuelle » et « référence virtuelle » dans la mesure où ce modèle met l'accent sur ce qui rapproche et maintient distincts, du point de vue de la référence, signification lexicale et référent discursif.

---

[1] Marandin 1992a, et précédemment chapitre II § II.2.1.
[2] Ou plutôt, s'il existe, le statut référentiel d'autres séquences, comme les séquences verbales, n'est pas de même nature que celui des séquences nominales, Milner 1976, p. 63.
[3] En reprenant la formalisation traditionnelle en linguistique issue de la théorie X-barre, établie par Chomsky. On ne présente pas ici les discussions dont a fait l'objet ce formalisme.
[4] Par exemple, Kleiber 1981, Tyvaert 1995.

Dans le cadre du modèle de Milner, une unité lexicale hors emploi de catégorie N, par exemple « homme », n'a qu'une « référence virtuelle », qui est sa signification. Un GN, une unité de discours, a, lui, une « référence actuelle », propre à construire des référents discursifs (par exemple, « un homme passe », « tout homme est mortel », etc.). La référence virtuelle, le sens d'une unité lexicale, « pèse » sur la construction du référent discursif[1] : c'est sur ce point que s'articulent langue et discours. En effet, ce n'est pas parce que la référence se construit *en* discours qu'elle est nécessairement déclenchée *par* le discours : on devrait sinon pouvoir construire n'importe quel objet de discours à partir de n'importe quel mot.

Observant au contraire que « n'importe quelle séquence nominale n'est pas associée à n'importe quel segment de réalité », Milner[2] relie la question du rapport sens/référence à celle de la propriété de discrimination des unités lexicales[3] : les unités lexicales se distingueraient entre elles notamment en fonction du type de « segment de réalité » qu'elles pourraient désigner. Toutes les unités lexicales ne sont donc pas équivalentes du point de vue du référent qu'elles permettent de construire. Le mécanisme est décrit ainsi par Milner : « Une unité lexicale étant choisie, certains segments sont d'emblée éliminés en tant que références possibles ; en ce sens, à chaque unité lexicale individuelle, est attaché un ensemble de conditions que doit satisfaire un segment de réalité pour pouvoir être la référence où interviendrait crucialement l'unité lexicale en cause. Cet ensemble de conditions décrit donc un *type* (ou si l'on veut une *classe*) de référence possible ; il est distinct des segments de réalité, mais pèse sur eux. Pour exprimer cette situation, on pourrait recourir aux termes suivants : le segment de réalité associé à une séquence est sa *référence actuelle* ; l'ensemble de conditions caractérisant une unité lexicale est sa *référence virtuelle*.[4] »

Notons qu'en dépit de formulations (telles que « segment de réalité ») qui ont pu faire dire que Milner, malgré ses ambitions, n'échappait pas au « problème ontologique des universaux », cette proposition en termes de références actuelle et virtuelle reste dans un cadre non réaliste. En effet, la notion de « classe » ou de « type » est chez Milner de nature artefactuelle ; une classe n'est pas constituée à partir des éléments du « réel », seul le discours peut l'actualiser, et ce n'est plus alors le même type d'unité qui est mis en cause : « Une unité lexicale ne peut avoir de référence actuelle que si elle est employée [...]. Mais, d'autre part, si l'on considère les emplois eux-mêmes, ce ne sont pas aux unités lexicales comme telles que sont associés des segments de réalité, mais bien aux *groupes nominaux* pris dans leur ensemble.[5] »

L'« atome référentiel, c'est donc le groupe nominal » et non le nom, le N, qui en constitue la tête.

Milner propose d'observer la différence entre référence actuelle et référence virtuelle à partir de différents types de reprises anaphoriques. Quand l'anaphore reprend un N (une référence virtuelle), l'anaphorique ne construit pas les mêmes

---

[1] Dans le cadre d'un autre modèle, Berrendonner a pu soutenir de semblables propositions : « Grâce aux référents [entendus ici comme « syntagmes nominaux définis »], la langue contient la mention de *certaines* de ses conditions d'emploi », Berrendonner 1978, p. 47.
[2] Milner 1976, p. 63.
[3] Milner parlera plus tard [1989] de « facteurs individuation lexicale » et proposera un modèle de l'individuation lexicale à trois composantes, Milner 1989, p. 324 et suiv.
[4] Milner 1976, p. 64.
[5] *Id.* (c'est nous qui soulignons).

objets de discours que ceux constitués par la source de l'anaphore ((1) ci-dessous) ; quand l'anaphorique reprend un GN (une référence actuelle), c'est le même référent discursif qui est visé (2) :

(1) J'ai vu dix lions et toi tu *en* as vu quinze

« en » reprend ici un N « lions », c'est-à-dire une référence virtuelle : il ne s'agit pas des mêmes référents visés (il ne s'agit pas des mêmes lions) ;

(2) J'ai capturé dix des lions et toi tu *en* a capturé quinze

« en » reprend ici un GN « les lions », c'est-à-dire une référence actuelle : il s'agit des mêmes référents (c'est le même ensemble de lions qui est visé).[1]

C'est cet effort d'articulation et de distinction entre construction du sens et construction de la référence qui caractérise, nous semble-t-il, l'apport de Milner. En effet, comme le signale d'ailleurs Milner lui-même[2], la dualité qu'il propose s'apparente à celle précédemment proposée par Frege. Mais si Frege distingue d'une part la « Bedeutung » (dénotation) et le « Sinn » (sens), c'est pour les opposer[3], alors que Milner s'attache à montrer comment les deux se déterminent l'un par rapport à l'autre ; c'est pourquoi, précise-t-il, « il semble préférable d'utiliser une terminologie qui ne dissimule en rien l'articulation[4] ».

L'autre apport de Milner consiste à spécifier la « plasticité[5] » particulière des GN en discours, qui leur permet de désigner différents types d'objet : « Bien qu'un nom ordinaire puisse désigner des individus totalement distincts suivant les énoncés, il reste toujours possible de définir de manière générale la classe des êtres dont ce nom est la désignation et inversement d'exclure *a priori* des êtres qui ne pourront jamais être désignés par lui.[6] »

Milner pose ce faisant la possibilité de définir la référence d'un nom hors contexte : c'est ce qu'il nomme son « autonomie référentielle », qui correspond approximativement à la notion logique de classe[7]. Son propos vise ici à distinguer la relation référentielle établie par le nom et celle établie par le pronom, soit les interprétations de (4) et de (5) ci-dessous :

(4) le livre est beau

(5) il est beau

Dans (4), l'interprétation est possible « même si l'on ignore de quel livre il s'agit. [...] Inversement, l'énoncé (5) n'est, dans les mêmes conditions, absolument pas

---

[1] Milner 1976, p. 64.
[2] Milner 1989, p. 341.
[3] C'est l'exemple bien connu des deux expressions « étoile du matin » et « étoile du soir » qui, tout en ayant un sens différent, désignent le même référent. Frege 1971 [1892], p. 102-126.
[4] Milner 1978, p. 26.
[5] Nous revenons dans le chapitre V sur cette notion de plasticité.
[6] Milner 1978, p. 198-199.
[7] Voir aussi sur ce point le chapitre V.

interprétable[1] ». La différence tient à ce qu'en (4), la référence virtuelle de *livre* est déterminée hors emploi : à ce titre *livre* est dit référentiellement autonome[2]. L'unité *livre* correspond à une classe qui circonscrit le type d'objet singulier auquel il est possible de référer en discours. Par opposition, « sera donc non autonome une unité dont la référence virtuelle ne peut être définie sans mentionner l'unité elle-même en tant qu'elle est énoncée dans un énoncé singulier[3] ». Le cas typique des unités non autonomes est représenté par les pronoms. Ainsi la référence virtuelle du pronom *je* ne peut être autonomisée que par rapport à son emploi dans un énoncé, elle est suspendue à l'énonciation ; il y a, dans ce cas, circularité de la référence virtuelle : la propriété définissante du référent est elle-même suspendue à l'énonciation[4].

Il importe de noter que l'autonomie référentielle dont parle ici Milner porte sur *le rapport entre l'unité lexicale et l'énoncé où elle est insérée*, et non sur le rapport entre l'unité et le référent qu'elle désigne : on ne sort donc pas du cadre non réaliste et cette hypothèse ne contredit pas celle « de la construction des objets de discours[5] ».

Les propositions de Milner articulant sens et référence par le biais des notions de référence virtuelle et référence actuelle présentent l'avantage, nous semble-t-il, de pouvoir circonscrire le type d'unité linguistique minimale permettant de référer[6] : c'est une molécule syntaxique de type GN, construite à partir d'un atome syntaxique de type N.

Une fois ces distinctions établies entre unité lexicale et groupe nominal, comment peut-on penser, de façon plus précise, l'inscription de la référence dans une unité lexicale ? En quoi peut consister la référence virtuelle d'une unité lexicale hors emploi ?

Comme précédemment dans le chapitre II, nous emprunterons à D. Corbin[7] les éléments de sa démonstration concernant le cas particulier des mots construits[8] ; ceux-ci présentent en effet l'avantage de faire « voir » le rapport entre sens et référence : « Les sens des mots construits sont des exemples privilégiés pour observer les relations entre le sens linguistique et les catégories référentielles : dans

---

[1] Milner 1978, p. 199.

[2] Une unité est dite référentiellement autonome quand « les conditions de possibilité de désignation sont indépendantes de l'énoncé particulier où l'unité est employée : l'unité ne tire sa capacité référentielle que d'elle-même », *Ibid.*, p. 333.

[3] Milner 1976, p. 65.

[4] Milner 1978, p. 333-334.

[5] Kleiber reconnaît sur ce point que la solution de Milner reste satisfaisante dans la mesure où elle « montre qu'un item lexical, tout en n'entretenant pas de relation directe avec des êtres ou objets précis de la réalité, est malgré tout en "prise" avec la référence par le biais de conditions d'application référentielle. En second lieu, elle ne nécessite nul engagement ontologique, puisqu'en parlant de "segments de réalité", elle ne place pas en première ligne la question de l'existence des référents », Kleiber 1981, p. 20.

[6] Dans un autre cadre, qui se fonde sur des arguments de nature logique et de nature linguistique, Michel Le Guern propose une description du groupe nominal qui permet de le faire voir également comme l'unité référentielle minimale ; nous revenons sur cette description dans le chapitre V.

[7] Bien que Corbin tienne à marquer ses distances avec le modèle proposé par Milner, voir par exemple Corbin et Temple 1994, p. 91, n. 4.

[8] « La spécificité d'un mot construit par rapport à un mot non construit est que l'interprétation sémantique du premier est compositionnelle par rapport à sa structure interne », Corbin 1990, p. 176. Ainsi si « maisonnette » est un mot construit, « omelette » n'en est pas un (« omel- » ne constitue pas une base suffixable en français).

la mesure où ces mots sont construits par des opérations linguistiques, leur sens peut être calculé de façon proprement linguistique, indépendamment des catégories que ces mots dénomment.[1] »

II.2.3 - CONSTRUCTION DE LA RÉFÉRENCE SUR LA BASE DE LA SIGNIFICATION LEXICALE

Rappelons que la problématique du rapport sens/référence repose sur deux principales questions :

- comment expliquer que, pour une forme lexicale donnée, il y ait stabilité de la signification lexicale d'une part et instabilité des référents discursifs d'autre part ?

- comment penser un modèle de construction de la référence sur la base de la signification lexicale qui inclut l'activité des sujets parlants (les pratiques discursives, les stratégies de désignation, etc.) ?

Le modèle de D. Corbin propose un cadre de réponse à ces deux questions. En effet, la base de son programme de recherche[2], mené dans le domaine de la morphologie dérivationnelle, consiste à rendre compte à la fois des unités attestées (constituant le « lexique conventionnel », illustré par le dictionnaire) et des unités possibles (constituant le « lexique dérivationnel », produisant des unités dont l'usage ne s'est pas emparé, ou pas encore emparé).

Sa position suppose que c'est la langue qui régit, ou programme, les désignations référentielles, dont certaines peuvent être jugées, pour des raisons sociales ou culturelles, « inacceptables » et donc n'être jamais employées (du moins en synchronie).

Ainsi, dans son modèle :

- la signification lexicale d'une unité peut permettre la construction de référents de nature hétérogène, l'hétérogénéité étant elle-même « programmée » ; autrement dit, l'hypothèse d'un conditionnement linguistique n'oblige en rien à penser un rapport univoque et constant entre sens et référent : c'est l'exemple donné ci-dessous de l'analyse du mot « chinois » ;

- le rôle des pratiques sociales consiste moins, dans ce cadre, à créer la référence qu'à la révéler : cette hypothèse permet donc d'intégrer l'influence des sujets parlants dans la fixation de la référence. Le lien sens/référence est maintenu dans son « indétermination », laissant une latitude de décodage ; c'est l'exemple donné ci-dessous de l'analyse du mot « fenouillette ».

Ce modèle qui défend le rôle du sens lexical dans la construction de la référence n'enlève donc rien à la contingence observable et observée des objets de discours ; il lui donne au contraire un cadre qui la révèle.

---

[1] Corbin et Temple 1994, p. 6.
[2] Corbin 1987.

On reprend ci-dessous, de façon schématique, les différents niveaux que pose Corbin pour penser l'articulation entre sens et référence (A). On présente ensuite deux exemples d'analyse qui illustrent chacun un point particulier : le premier, l'analyse de « chinois », montre comment le même sens linguistique peut construire des référents distincts (B) ; le second, l'analyse de « fenouillette », met en valeur le rôle des pratiques sociales dans la construction des référents (C).

A - Distinction des niveaux dans le modèle de Corbin

Notons d'emblée que, si, comme Milner, Corbin considère qu'il y a distorsion entre sens et référence, elle ne postule pas, comme lui, un rapport direct entre les deux. Elle pose en effet un niveau intermédiaire entre ce qu'elle nomme les « catégories sémantiques » (définies en intension) et les « catégories référentielles[1] » (définies en extension) : c'est le niveau des « catégories pré-référentielles », définies en intension et en extension, qui « représentent le résultat des découpages conceptuels que les propriétés sémantiques permettent d'opérer et de dénommer ». C'est seulement lorsqu'elles sont « adaptées en fonction de notre appréhension pragmatique et culturelle des choses » qu'elles deviennent des catégories référentielles. Autrement dit, le modèle est prêt à accueillir des unités bien construites du point de vue de la langue mais bloquées au niveau pré-référentiel pour des raisons non linguistiques mais culturelles ou sociales.

C'est dans le cadre de ces distinctions que Corbin démontre, de façon parallèle, d'une part l'autonomie du sens lexical par rapport à la référence et d'autre part le conditionnement sémantique de la référence : « Défendre l'idée que les catégories sémantiques ne sont pas nécessairement isomorphes aux catégories référentielles et que ne leur correspondent pas nécessairement des catégories référentielles présuppose qu'il existe des catégories sémantiques associées aux mots, donc un sens lexical, [qui] détermine et conditionne, conjointement à d'autres facteurs, leur interprétation contextuelle.[2] »

B - L'analyse du mot « chinois » : sens unique et multiplicité référentielle

Le nom « chinois » est un mot construit par conversion à partir d'un adjectif[3] : « ce procédé de construction de mots illustre l'hypothèse qu'à des catégories sémantiques homogènes peuvent correspondre des catégories référentielles conçues comme hétérogènes[4] ». On constate en effet que le mot « chinois » renvoie à des catégories référentielles différentes : personne ; langue ; petite orange amère ; passoire.

Corbin et Temple proposent de voir dans le cas des noms construits par conversion à partir d'adjectifs des cas de polyréférence, à rapporter non à des sens différents

---

[1] Corbin et Temple sont plus précises : les catégories référentielles « forment un sous-ensemble des catégories conceptuelles dénommables par les unités lexicales, c'est-à-dire celles qui sont nommables de façon fixe et codée par une expression linguistique », Corbin et Temple 1994, p. 6.
[2] *Ibid.*, p. 7.
[3] Corbin et Temple 1994, p. 12-13, qui analysent, dans les mêmes termes, les mots « simple » et « bleu ».
[4] *Ibid.*, p. 13.

mais à des sens uniques, à des principes organisateurs uniques subsumant les différentes catégories référentielles[1].

Le point de la démonstration constitue un article en soi : nous ne rentrerons pas dans le détail de l'analyse qui nécessiterait en outre d'exposer l'appareillage descriptif qui constitue le modèle d'analyse de Corbin [1987]. De façon schématique, on peut dire que le sens unique des noms construits par conversion d'adjectifs repose sur le sens de la base lexicale à l'origine de la conversion : la propriété sémantique retenue de l'adjectif au terme de la conversion constitue un « type », une propriété saillante, qui doit servir de propriété dénominative aux entités susceptibles d'être dénommées par le mot construit par conversion de l'adjectif.

C'est sur la base d'une propriété saillante de l'adjectif que le mot construit peut renvoyer à des catégories référentielles hétérogènes.

C - L'analyse du mot « fenouillette » : intervention des pratiques sociales dans la fixation de la référence

Cet exemple illustre le cas où le sens d'un mot construit (les catégories sémantiques qu'il construit) ne peut pas servir de support à la dénomination de catégories référentielles, ceci pour des raisons sociales et culturelles ; il faut, pour qu'elle puisse être employée, qu'une dénomination respecte nos modes d'organisation conceptuels.

La démonstration de ce point étant assez complexe, nous n'en reprenons que les grands traits. Corbin et Temple étudient le cas de « fenouillette[2] », dont la définition usuelle est : « petite pomme grise dont le parfum rappelle celui du fenouil ». Il s'agit d'expliquer le rapport entre la base du mot construit (fenouil) et le référent désigné (pomme), rapport visiblement incongru puisqu'« il ne semble pas possible en français de dénommer un végétal comestible par le nom d'un autre végétal comestible », à moins de construire une expansion, comme on le voit dans les unités de type « poire *d'avocat* » (fruit de l'avocatier) ou « poire *de terre* » (nom régional du topinambour). Cette impossibilité est à l'origine du report de la signification de « fenouillette » sur l'« *odeur* de fenouil » : c'est une propriété stéréotypique de « fenouil » qui est alors sélectionnée. Or la catégorie référentielle des « odeurs » n'existe pas en français : « on ne trouve pas dans le lexique français de dénominations de catégories rassemblant différents objets ayant le même parfum ». Dès lors, pour expliquer l'acception usuelle de « fenouillette », il faut supposer qu'à une catégorie sémantique (ici « odeur ») ne correspond pas nécessairement une catégorie référentielle du même type : « Il existe des catégories sémantiques dont certaines propriétés ne peuvent servir de support à la dénomination de concepts. [...] La base *fenouil* sur laquelle s'applique *-ette* pour construire *fenouillette* n'est pas un mot, c'est une forme donnant corps à l'une des propriétés sémantiques du sens métaphorique de fenouil – celle qui renvoie au parfum –, propriété qui ne peut pas servir à organiser une catégorie référentielle.[3] »

---

[1] Corbin et Temple 1994, p. 13. Il y a, dans ces cas, une catégorie préréférentielle unique et des catégories référentielles différentes : « Les catégories préférentielles correspondant aux catégories sémantiques définies par les sens des mots construits ne sont pas isomorphes aux catégories référentielles dénommées par les mots ».
[2] *Ibid.*, p. 17-24.
[3] *Ibid.*, p. 24.

L'une des conclusions que permet de formuler ce traitement de « fenouillette » est que « bien que les mots aient un sens et que ce sens permette de référer, les mots ne sont pas directement des dénominations[1] » : ils ne le sont que dans la mesure où ils respectent nos modes d'organisation du monde.

Les exemples d'analyse de « chinois » et de « fenouillette » proposés par Corbin montrent que la construction de la référence est toujours doublement configurée : et par le sens lexical et par les contingences discursives. Compte tenu de cette dualité, le sens est nécessairement sous-déterminé par rapport à la référence[2]. C'est cette sous-détermination qui constitue l'espace de « jeu », interprétatif et désignatif, qui revient au sujet parlant. La notion de sous-détermination montre que l'hypothèse de l'instabilité des objets de discours peut être compatible avec celle de la contrainte sémantique sur la construction de la référence.

Chacun dans leur approche, Milner et Corbin s'attachent à faire apparaître à la fois une disjonction nette entre sens et référence et une contrainte sémantique sur la construction de la référence. En ce sens, le rapport entre sens et référence se donne comme un rapport de sous-détermination : la langue et la référence n'étant pas coïncidentes, la construction de la référence n'est qu'en partie réalisée par la langue.

II.3 - Conclusions et mise en perspective

Dans le paragraphe I de ce chapitre, nous avions tenté de dégager, au-delà des présupposés réalistes du discours classique, les zones de manifestation de la référence en indexation : à ce titre, le document, en tant qu'élément constitutif de l'« extérieur » de l'indexation, nous est apparu comme un objet construit, distinct de la « réalité éditoriale ». Parallèlement, nous nous étions interrogée sur la stabilité référentielle attachée, dans le discours classique, aux unités d'indexation : hors du modèle réaliste, ce type de stabilité référentielle nous a paru relever davantage d'un principe de réception que d'un principe de production.

Afin de pouvoir clarifier ces aspects de la référence en indexation, nous avons cherché à nous dégager du modèle réaliste. Nous avons alors sollicité le modèle de la référence sur lequel s'appuie l'approche linguistique. Dans ce cadre, est apparue une série de distinctions :

(i) entre éléments de la référence : l'acte de référence doit être distingué des propriétés référentielles des expressions linguistiques ;

(ii) entre types de référent : l'approche linguistique établit une distinction entre référent mondain et référent discursif ;

(iii) entre expressions linguistiques : seul le GN est susceptible de construire un objet de discours, mais cette construction passe par l'exploitation de la signification de l'unité lexicale de catégorie N qui en constitue la tête ;

(iv) entre sens et référence : tout en restant distincts, sens et référence établissent entre eux un rapport de sous-détermination ;

---

[1] Corbin et Temple 1994, p. 25.
[2] *Ibid.*, p. 17.

(v) entre langue et usage : la référence, potentiellement déterminée en langue, se fixe par l'usage, par l'intervention des sujets parlants.

L'établissement de ces distinctions, propres à une approche non réaliste de la référence, permet, « rétroactivement », de préciser ce qui, dans le discours classique sur l'indexation, relève du modèle réaliste et ce qui, par conséquent, ne pourra apparaître dans le cadre du nouveau modèle de la référence en indexation proposé ci-après.

Ainsi, la distinction (iv) contrecarre l'hypothèse du discours classique concernant la correspondance entre stabilité référentielle et stabilité linguistique en indexation : ce n'est pas parce que l'on emploie le même mot que l'on désigne la même chose, même si ce mot est un « descripteur ». L'approche linguistique de la référence montre en effet qu'un accord de désignation ne suppose pas un « accord » de signification[1] : le sens d'une unité lexicale s'établit bien en deçà de l'intervention des sujets parlants et reste sous-déterminé.

Si tous les mots ne découpent pas la réalité de la même façon, tous les mots ne se valent pas. À ce titre, la notion du descripteur comme « terme préférentiel » (comme terme retenu de préférence à d'autres tenus pour équivalents) devient problématique. Nous sommes amenée à nous interroger sur l'aspect de la sélection des termes en indexation : si un objet de discours n'est construit qu'au travers de saisies multiples opérées grâce aux sens différents des unités lexicales, l'indexation doit-elle restreindre le nombre des accès lexicaux aux documents, c'est-à-dire restreindre les possibilités de saisie multiple ?

La distinction (iii) souligne la différence entre N et GN du point de vue de leurs propriétés référentielles : un N, une unité lexicale hors emploi, ne peut que désigner une classe de référents possibles ; un GN, l'unité lexicale en discours, désigne, lui, un élément singulier de cette classe. C'est lui qui peut établir une relation référentielle stable ; or les descripteurs se donnent généralement sous la forme linguistique de N.

Le point (v) fait apparaître la dimension du discours en indexation : elle est complètement absente des approches normatives, mais se dégage nettement sous l'angle de l'approche linguistique de la référence. En effet, comme le rappellent Apothéloz et Reichler-Béguelin, « le problème du choix des dénominations ne doit pas être pensé dans le rapport entre la langue et le monde mais à l'intérieur du discours[2] ». La reformulation de la référence en indexation impose donc de circonscrire cette dimension discursive en indexation.

Si les points (iii), (iv) et (v), dégagés d'une approche linguistique de la référence, nous engagent à introduire la dimension du discours en indexation et à redéfinir la morphologie du descripteur[3], les points (i) et (ii) nous permettent d'ores et déjà d'approcher ce que peut être la construction de la référence en indexation.

---

[1] Tyvaert 1994, p. 48.
[2] Apothéloz et Reichler-Béguelin 1995, p. 266.
[3] Voir les chapitres IV et V.

III - La construction de la référence en indexation

Dans ce paragraphe, nous tenterons de décrire le processus de l'indexation en l'inscrivant dans le cadre du modèle d'approche linguistique de la référence (approche non réaliste) tel que nous l'avons précédemment décrit.

L'adoption de ce cadre nous conduit à distinguer :
- différents aspects de la référence : l'acte référentiel et les propriétés référentielles des expressions linguistiques utilisées en indexation ;
- différentes dimensions du référent, notamment les dimensions mondaine et discursive.

Pour penser ces différents aspects de la référence, il importe de préciser d'abord l'acte de référence effectué en indexation, duquel dépendent les propriétés référentielles des expressions linguistiques utilisées en indexation.

L'acte de référence est, en indexation, double :

- du côté de l'indexeur, il s'établit de la réalité éditoriale à la réalité documentaire. Le monde de référence des indexeurs n'est pas, pour reprendre les termes de Ricœur[1], le « monde ambiant » mais le « quasi-monde des textes ». L'acte de référence consiste alors en un *acte de discrétisation* sur ce monde des textes, qui permet de construire un univers proprement documentaire : l'univers des documents. Cet acte de discrétisation, s'il peut se traduire en « mots », par l'assignation de descripteurs, ne suppose pas en soi une nomination à proprement parler : la dénomination peut n'intervenir qu'après, si l'on veut indiquer expressément la base sur laquelle des documents ont été rapprochés. À ce niveau, l'indexeur cherche à donner accès à des *classes de référents discursifs* et non à des objets de discours particuliers à ces classes. Il utilisera pour ce faire, des unités lexicales hors emploi, pourvue d'une référence virtuelle ;

- du côté des utilisateurs, l'acte de référence s'établit de la réalité documentaire à la réalité éditoriale, sur la base de leurs représentations de la réalité mondaine : ce sont elles qui guident la recherche de l'utilisateur à travers l'univers documentaire. L'acte de référence consiste ici en un *acte de stabilisation du monde de référence* : il s'agit de construire, à travers l'exploration de l'espace documentaire, des objets de discours perçus comme stables (*i.e.* « parlant de la même chose »). Les référents discursifs n'intéressent les utilisateurs que par les représentations des référents mondains auxquels ils renvoient : l'utilisateur cherche à capter des représentations discursives du monde. Pour cela, il opère différentes saisies sur le monde des textes lui permettant de construire de proche en proche ses objets de discours : l'utilisateur traverse les différentes classes de documents constituées par le biais d'un même individu linguistique, *un groupe nominal*, qui peut alors apparaître comme un thème de discours.

La proposition de ce modèle de la référence documentaire nous écarte de façon radicale des approches classiques de l'indexation et du descripteur :

---

[1] Ricœur 1986, p. 141.

- l'indexation ne peut plus être tenue pour une opération symétrique : les objets construits par les indexeurs (des classes de discours) ne correspondent pas à ceux construits par les utilisateurs (des objets de discours) ;

- le descripteur ne peut donc plus être conçu de façon homogène. La référence du descripteur apparaît en effet sous une double forme : le descripteur des indexeurs pointe sur une source/des sources *via* un document. Le descripteur des utilisateurs pointe sur un objet mondain *via* des objets de discours. Dans les deux cas, la référence d'un descripteur est à chercher, non dans le « mot » lui-même, mais dans les discours, les textes, auxquels il donne accès.

L'approche documentaire de la référence conduit à postuler l'existence d'un lien entre référent mondain et référent discursif, lien qui repose sur une croyance, celle de partager un univers réel commun[1]. À ce titre, l'indexation peut être considérée comme l'opération qui crée cet « univers réel commun », ou du moins perçu comme tel. En effet, l'indexation, appréhendée sous l'angle de la référenciation, doit permettre d'établir une correspondance entre les différents espaces de représentations, celui des auteurs, celui des indexeurs et celui des utilisateurs : les sources à partir desquelles les indexeurs construisent des documents constituent des représentations, des saisies de la réalité mondaine ; ces saisies sont ressaisies par les indexeurs qui construisent des documents ; ces documents sont à nouveau ressaisis par les utilisateurs qui construisent leurs objets de discours, autres représentations de la réalité mondaine.

L'indexation ne se justifie que dans la mesure où elle peut assurer une certaine continuité entre ces différentes formes de représentations. Nous essayerons de dégager, dans le chapitre IV, des stratégies d'exposition propres à l'indexation permettant de créer un « espace commun » de représentations.

Auparavant, nous nous proposons ci-dessous de préciser comment l'indexation peut se comprendre comme un acte de référenciation, c'est-à-dire comme un acte établissant un *double processus de discrétisation et de stabilisation.* C'est sous ces deux rapports que l'on étudiera :

(i) d'une part, la construction des référents de l'indexation, c'est-à-dire le passage du texte (ou source) au document : il s'agit là d'une étude du processus de discrétisation qui relève de l'acte de référence réalisé par les indexeurs ;

(ii) d'autre part, la construction de l'effet de stabilité référentielle, c'est-à-dire l'utilisation documentaire du « désignateur rigide ». Il s'agit là d'une étude de la stabilisation référentielle (première approche[2]), qui porte non plus sur l'acte de référence mais sur les propriétés référentielles des expressions linguistiques utilisées en indexation.

---

[1] Michel Le Guern (communication personnelle) : « Que la référence porte en fait sur des objets de discours, c'est vrai ; mais l'indexation n'existerait pas s'il n'y avait pas dans l'esprit des indexeurs et des utilisateurs la croyance en un univers commun réel, les objets de cet univers, que ce soient des objets concrets ou des abstractions, étant identifiés de manière plus ou moins floue avec les objets de discours ».

[2] L'étude sera approfondie dans le chapitre V.

III.1 - Construction du document en indexation

Nous souhaitons ici étayer l'hypothèse que, en tant que processus de construction de la référence, l'indexation construit son référent (le document), et intègre, ce faisant, la procédure de sélection documentaire au cœur même de son mécanisme de fonctionnement.

Cette hypothèse a été formulée sur la base des résultats de l'expérimentation de « mise en document » que nous avons menée[1] : objet textuel et objet documentaire ne sont pas apparus, dans tous les cas, isomorphes. Ce décalage peut être appréhendé de deux façons :

- dans les termes de l'approche « instrumentale » : la sélection documentaire n'est que le résultat de contraintes liées aux besoins des utilisateurs ;

- dans les termes de l'approche « procédurale » : la sélection documentaire se comprend comme un principe de fonctionnement documentaire (du document). La transformation d'un objet en document constitue le propre de l'indexation, entendue comme mise à disposition du savoir (ou plutôt mise à « construction » du savoir). C'est là une piste de recherche proposée par Escarpit[2] et que nous retenons dans cette étude.

Pour caractériser la construction du document à partir de sources, nous invoquerons des mécanismes de transformation permettant, d'une part, de distinguer ce qui change dans le passage de la source au document (le contexte*, selon notre hypothèse) et, d'autre part, d'indiquer le rôle du document par rapport à la source (rôle d'interprétant, selon nous[3]).

Pour étudier ces deux aspects, nous adopterons une approche par point de vue :

- nous étudierons d'abord la construction du document du point de vue de la source (III.1.1) ; pour montrer que c'est le contexte de la source qui est modifié, nous utiliserons des éléments descriptifs issus de la théorie de l'énonciation[4] ;

- nous étudierons ensuite la construction du document vue du côté du document (III.1.2) : pour caractériser le rôle que joue le document par rapport à la source, nous utiliserons, là encore, une théorie, celle de Peirce, et plus particulièrement la notion d'interprétant qu'il propose.

---

[1] Voir, dans ce chapitre, § I.3.1.

[2] « Ce qui est certain, c'est que le pas décisif a été franchi lorsque l'homme a institué le *document*, cumulation de traces fixes et permanentes [...] où les réponses données, en feedback, à travers le temps, aux expériences antérieures restent disponibles pour une lecture, c'est-à-dire pour l'exploration libre de toute contrainte événementielle ou chronologique, en fonction du projet et de la stratégie destinée à le réaliser. En d'autres termes : il y a constitution d'un savoir », Escarpit 1991, p. 62-63.

[3] Nous remercions Michel Le Guern pour nous avoir guidée sur cette voie.

[4] Suivant en cela aussi une proposition de Michel Le Guern : qu'il en soit ici remercié. L'énonciation sera ici envisagée comme l'étude des « allusions qu'un énoncé fait à l'énonciation, allusions qui font partie du sens même des énoncés », Ducrot et Schaeffer 1995, p. 603.

Sans le recours à ces emprunts théoriques, les processus à l'œuvre dans la construction du document en indexation restent peu visibles.

Enfin, en III.1.3, nous proposerons une schématisation du passage de la source au document.

III.1.1 - LA CONSTRUCTION DU DOCUMENT VUE DU CÔTÉ DE LA SOURCE

A - Conjecture : la source vue comme une énonciation

En première approche, on peut définir l'énonciation[1] comme « l'événement historique constitué par le fait qu'un énoncé a été produit[2] » ; en cela, l'énonciation suppose un énonciateur et une situation d'énonciation. L'énoncé, en tant que produit de l'énonciation, porte, en lui, sous forme de « traces », les marques de la « subjectivité » de l'énonciation[3] : c'est à ce titre qu'un énoncé peut être interprétable, même coupé de ses instances de production initiales, mais aussi réinterprétable en fonction des nouvelles situations de discours dans lesquelles il apparaît.

Ce cadre rapidement posé, on propose de voir la source comme une « énonciation »* et le document comme un « énoncé »* dont, par ailleurs, certains traits énonciatifs, tels que l'énonciateur (l'auteur d'un article par exemple) et le contexte de production (le nom du journal d'où est issu un article par exemple), peuvent être notifiés ailleurs dans le système d'information[4].

La « mise en document » aurait donc pour effet d'extraire, en partie, l'énoncé de sa situation d'énonciation, pour le réintroduire dans un autre espace d'interprétation (une collection documentaire). La mise en document opère donc un changement de contexte, qui modifie les caractéristiques de la source : la source devenant autonome par rapport à son utilisation initiale est ainsi susceptible de connaître d'autres usages que ceux pour lesquels elle avait été conçue[5].

Escarpit procède à une analyse de la constitution du document dans des termes semblables[6], mise à part que, dans son cadre, la notion de source est approchée en termes d'« événement », qu'il s'agit de stabiliser par une transformation en document[7].

---

[1] Ici entendue sous l'angle de l'acte d'énonciation, et non plus sous l'angle de la théorie de l'énonciation, comme précédemment.
[2] Ducrot et Schaeffer 1995, p. 603.
[3] Kerbrat-Orecchioni 1980.
[4] Nous reprenons ce point ci-après dans le chapitre IV.
[5] Il importe sur ce point de relever qu'il n'existe aucun texte, aucune photographie, aucun disque, etc., spécifiquement dédiés à l'indexation ou à la recherche documentaires : c'est un *usage documentaire* des sources qui est créé. Voir, sur ce point, Varet 1995 : « Le livre *devient* un document lorsqu'il est invoqué comme texte à l'appui ».
[6] Escarpit 1991, chap. 8, p. 121-147.
[7] D'autres auteurs insistent sur cet aspect du document comme « stable » ; ainsi Turner [1994], par exemple : la caractéristique principale du document est, selon lui, « d'avoir une certaine immuabilité lui permettant de traverser l'espace et le temps ». Dans le cadre de notre approche, cette caractéristique du document peut s'analyser en termes d'*effet* stabilisateur réalisé par la pratique d'indexation. Voir aussi sur ce point Escarpit [1991, p. 125] qui montre qu'un document n'est pas, à proprement parler, stable : « Un objet stable reste

Comme le fait remarquer Escarpit, introduire la dimension de l'énonciation dans l'approche du document permet de mettre l'accent sur la source de l'information[1], et non plus, comme dans les modèles mécanistes, uniquement sur le canal. Sur ce point, Escarpit propose la notion de « semi-document » pour capter les cas où la sélection documentaire n'opère que sur le canal : « En fait, la machine M3[2] n'est qu'un simple codeur-décodeur du temps. Elle n'agit ni au niveau de la source ni au niveau du destinataire. Elle agit au niveau du canal. Elle se contente de coder en synchronie documentaire la diachronie événementielle. Ce qu'elle restitue est simplement une image répétitive (sonore ou visuelle) de l'événement, soumise à la loi de dégradation de l'information. [...] Ce n'est donc qu'un semi-document.[3] »

Avec la mise en valeur de cette notion de « semi-document » apparaît en propre la spécificité du document : le document est ce qui permet de lire une source non plus comme « la réactivation de l'événement »[4] (niveau du canal, codage/décodage) mais comme la « production d'une information nouvelle[5] ». La source est *disponible* pour de nouvelles lectures et de nouveaux usages[6].

Si l'on définit une source documentaire comme une énonciation susceptible d'être décontextualisée (transformée en énoncé), tous les objets du « quasi monde des textes » constituent-ils, au même titre, des sources documentaires ? De quels types de propriété doit être dotée une source documentaire ?

B - Propriétés de la source

Pour qu'elle puisse être transformée en document, une source doit, nous semble-t-il, être dotée de deux propriétés :
(i) une propriété d'autonomie par rapport à son contexte de production ;
(ii) une propriété d'usage : une source doit pouvoir être détournée de son usage initial.

(i) Propriété d'autonomie de la source

Il est constant de remarquer que « l'information documentaire ne représente qu'une partie de l'information en circulation, la partie qui peut être détachée de ses contraintes originelles[7] ».

L'information en circulation se distingue, selon nous, de l'information documentaire sur deux points : elle n'est autonome ni d'un point de vue physique (pas de support) ni d'un point de vue « logique » (pas de contexte de production

---

emporté par le temps, la stabilité n'étant que le non-changement des relations qui définissent sa configuration caractéristique (son *pattern*) aux yeux d'un observateur emporté par un temps qui est, par convention pratique, supposé le même ».
[1] Voir le schéma de la communication présenté dans le chapitre I, § I.2.
[2] La machine M3 correspond, dans la typologie d'Escarpit, à une « machine à mémoire » de type magnétophone ; dans le contexte, elle s'oppose à une machine de type M6, dite « machine à langage ». Escarpit 1991, p. 109.
[3] *Ibid.*, p. 126.
[4] *Ibid.*, p. 126.
[5] *Id.*
[6] *Id.* : « La *disponibilité* pour un balayage volontaire et non forcément synchrone, non plus au niveau du canal (codage-décodage) mais au niveau de la source et du destinataire qui en ont chacun de son côté l'initiative, est donc ce qui caractérise le document ».
[7] Salaün 1991, p. 139 et suiv.

explicite : on pourrait dire qu'en cela l'information en circulation reste « suspendue » à l'acte d'énonciation et qu'elle n'existe pas en dehors d'elle).

Escarpit insiste surtout, dans l'approche du document qu'il propose, sur l'autonomie physique propre au document et qui lui vient du fait qu'il constitue un écrit[1]. La dimension de l'écrit apparaît dans son modèle comme proprement définitoire du document : on comprend alors pourquoi le document doit toujours, en indexation, comporter un texte[2].

Mais l'autonomie physique ne nous semble pas être le seul aspect de l'autonomie que la source doit présenter. Les interrogations récentes des professionnels sur le caractère documentaire ou pas (« indexable » ou pas) des « flux d'informations » transmis, par écrit, sur le réseau Internet[3] mettent l'accent sur l'importance, dans la construction documentaire, de pouvoir circonscrire un ensemble de paramètres de production de l'information (lieu, temps, acteur, notamment). Ces flux d'information posent la question de ce que nous avons appelé l'« autonomie logique » d'une source : il paraît en effet que, pour être décontextualisée, une source doit bel et bien avoir été constituée dans un contexte à même de laisser une « trace » dans la source elle-même.

Sur ce point, l'analyse documentaire des images est éclairante[4] : autant une image fixe peut, moyennant l'attribution d'une légende par exemple, être constituée en document autonome. Autant une image mobile, extraite d'une séquence filmique, pourra être difficilement constituée en document : le texte d'accompagnement devrait alors fonctionner comme une description du contexte des autres images de la séquence ; ce n'est pas là son rôle en indexation. En effet, l'autonomie « logique » de la source, qui signifie que le document construit garde une trace du contexte de production initial, est ce qui permet de *contraindre* les détournements d'usage que le document doit nécessairement autoriser pour permettre la production de nouvelles « informations ».

(ii) Propriété d'usage de la source

Une source doit pouvoir être utilisée moins pour l'usage explicite pour lequel elle a été conçue que pour d'autres usages[5], dont tous ne peuvent être connus ou répertoriés : c'est ici l'espace d'inventivité de l'utilisateur qui ne concerne pas nécessairement l'indexeur. La mise en document apparaît alors comme ce qui doit circonscrire l'espace de détournement possible d'une source.

La propriété d'usage multiple d'une source peut être appréhendée en termes de rentabilité ; ainsi du document iconographique : « Si l'on pose le problème en termes de rentabilité, l'image rentable, c'est-à-dire celle qui serait susceptible d'être intégrée à un très grand nombre de discours, serait celle qui pourrait être facilement

---

[1] La « fonction documentaire » correspond, pour Escarpit, à la stabilisation d'un message « sur un support qui le rend indépendant du temps et synchroniquement disponible », Escarpit 1991, p. 124.
[2] Voir notre interrogation à la fin du chapitre I.
[3] Par exemple les échanges écrits dans une liste de discussion.
[4] On reprend ici l'analyse que propose Odile Le Guern [1989, p. 428].
[5] Par exemple, un brevet, dont la finalité première est de protéger un inventeur, peut être, documentairement utilisé, dans une base de brevets, pour permettre d'informer sur les découvertes récentes.

décontextualisée, qui pourrait apparaître comme autonome et indépendante par rapport au contexte que constitue l'organisation séquentielle qui l'intègre.[1] »

Toutes les sources ne sont pas au même titre rentables, c'est-à-dire décontextualisables et recontextualisables dans le cadre d'usages différents ; ainsi, pour reprendre la boutade d'U. Eco, seul Proust est capable de mener une lecture « psychédélique » de l'indicateur des chemins de fer[2].

En effet, il y a sans doute des types de source qui sont moins rentables que d'autres : là encore, toute une étude précise reste à mener qui indiquerait, notamment par l'établissement d'une typologie par « genres » de production, un « degré » de détournement possible.

Pour préciser ce sur quoi porte, en indexation, la mise en document, nous avons sollicité quelques-uns des éléments d'une théorie de l'énonciation : à travers ce modèle, on a proposé de considérer la mise en document comme une opération portant sur le contexte d'énonciation, ou encore le contexte de production. De là peuvent se formuler les propriétés discriminantes des sources documentaires. Comment se passe la construction du document, vue du côté du document lui-même ?

III.1.2 - LA CONSTRUCTION DU DOCUMENT VUE DU CÔTÉ DU DOCUMENT

Comme précédemment, nous partirons d'une conjecture (ici, le document vu sous l'angle de l'interprétrant) pour arriver à déterminer, ne serait-ce que partiellement, des propriétés, ici celles du document. Parallèlement, nous tenterons de préciser le rapport qu'établit le document, en tant qu'interprétant, avec la source : est-ce un rapport d'interprétation ? est-ce un rapport d'utilisation ?

A - Conjecture : le document comme interprétant de la source

Pour capter le rôle du document par rapport à la source, il paraît fructueux de recourir au modèle proposé par Peirce. En reprenant ses termes, on dira que le document est l'interprétant de la source, c'est-à-dire non l'interprète mais « le moyen que celui-ci utilise pour effectuer son interprétation[3] ». Il constitue en cela les conditions d'interprétation d'une source : on se souvient en effet que l'interprétant est « la règle qui permet au *representamen* de renvoyer à un objet[4] ».

On constate en effet souvent ce paradoxe qui consiste à indexer un document non pour retrouver le document lui-même mais les informations qu'il contient[5]. Les informations, tout comme le document, n'existant pas en soi, hors d'une lecture ou

---

[1] O. Le Guern 1989, p. 428.
[2] Eco 1985 [1979], p. 74 : « Proust pouvait lire l'horaire des chemins de fer et retrouver dans les noms des localités du Valois les échos doux et labyrinthiques du voyage de Nerval à la recherche de Sylvie. Mais il ne s'agissait pas d'interprétation de l'horaire, c'était l'une de ses utilisations légitimes, presque psychédélique ».
[3] Peirce *in* Everaert-Desmedt 1990, p. 40.
[4] Voir précédemment, chapitre II (§ III.2), la présentation du modèle sémiotique de Peirce.
[5] Cf. Dachelet 1990 : « On a renoncé à l'espoir de fournir à l'utilisateur LA réponse à LA question posée. Une réponse, c'est aujourd'hui un document ou un ensemble de documents que l'utilisateur estimera pertinent ».

d'un usage[1], l'indexation ne peut directement pointer sur elles ; en revanche l'indexation peut donner à interpréter les « données » d'une source en « informations » *via* un document.

Ainsi le référent de l'indexation est-il bien le document qui, en tant qu'interprétant, désigne[2] sa source : par le document, l'indexation crée un espace intermédiaire de représentations communes où les utilisateurs peuvent interpréter les sources et construire leurs objets de discours, nous y reviendrons dans le chapitre IV.

Si, comme dans le cas du semi-document, l'indexation portait directement sur la source, il n'y aurait pas de décontextualisation possible et donc la construction des objets de discours ne serait que la reconstruction des objets du discours du texte. Dans le cas de l'indexation du document, la situation est différente : l'information ne se conçoit plus comme « donnée puis comme ressource stockable puis consommable, [mais] comme processus[3] ». C'est pour cette raison que si les descripteurs sont des accès aux documents, ils sont des accès *à* l'information et non des accès d'information, c'est-à-dire informatifs par eux-mêmes. Interprétant et non interprète, l'opération de mise en document organise donc un accès sous-déterminé (d'un point de vue interprétatif) aux sources.

Sur ce point, il apparaît que les différents types de « mise en document » des sources[4] ne modifient pas nécessairement l'information proposée à l'usager : plusieurs conditions d'interprétation différentes peuvent en effet permettre le même type de lecture.

Cette approche du document comme interprétant de la source permet de souligner que les objets documentaires ne sont pas, en indexation, des objets « interprétés » : il semble plutôt que les sources sont présentées aux utilisateurs (les interprètes) comme interprétables dans un cadre, dans un contexte que le document précise. Ce peut être, par exemple, pour un document constitué de plusieurs sources, les sources elles-mêmes qui constituent les unes pour les autres leur propre contexte d'interprétation ; ce peut être aussi les descripteurs, qui peuvent permettre d'interpréter une source par référence aux autres sources (textes) dont le mot porte la trace.

Vu sous l'angle de l'interprétant, le document est appelé à être redéfini aussi bien dans sa « nature » que dans sa « fonction ». Du point de vue de sa « nature », le document se laisse voir comme un « énoncé », inséré dans de nouveaux contextes d'énonciation. Pour ce qui est de son rôle, le document en tant qu'interprétant d'une source, s'il en donne les conditions d'interprétation, peut aussi en fixer les règles d'utilisation : en effet, si le propre d'une source est d'être détournée, encore faut-il que ce détournement puisse être contrôlé. Mais comment ?

---

[1] Dachelet 1990 : « Une base de données ne fait que stocker des données, données dont c'est à l'utilisateur lui-même de décider lesquelles constituent les informations qu'il cherche ».
[2] Il nous semble en effet que l'interprétant est, dans le cas du document, de type indiciaire.
[3] Batime 1995. Cf. aussi Capurro 1992 : « Information is not the end product of a representation process, or something being transported from one mind to the other, or, finally, something separated from a capsule-like-subjectivity, but an existential dimension of our being-in-the-world-with-the-others ».
[4] Voir les résultats de notre enquête rapportés dans le § I.3.1 de ce chapitre.

Préciser le rôle que tient le document en tant qu'interprétant de la source conduit à interroger ce que l'indexation cherche à transmettre via un document : des conditions d'interprétation ou bien des modes d'utilisation des sources ?

B - Indexation : mode d'interprétation ou mode d'utilisation ?

Pour discuter la proposition suivante de Valéry : « il n'y a pas de vrai sens d'un texte », Eco[1] propose d'établir la distinction suivante entre les deux notions d'« interprétation » et d'« utilisation » :

- l'interprétation d'un texte consiste à « respecter le monde possible » décrit dans le texte ;
- l'utilisation d'un texte se marque par l'introduction dans le discours d'« informations extratextuelles.[2] »

Ce qui distingue essentiellement interprétation et utilisation[3], c'est le type de contexte, interne ou externe au texte, dans le cadre duquel on se propose de lire, ou dans le cadre duquel on donne à lire, un texte : en effet, « un texte n'est pas autre chose que la stratégie qui constitue l'univers de ses interprétations – sinon légitimes – du moins légitimables. Toute autre décision d'utiliser librement un texte correspond à une décision d'*élargir l'univers de discours*. La dynamique de la sémiosis illimitée ne l'interdit pas, au contraire, elle l'encourage. Mais il faut savoir ce que l'on veut : faire subir un entraînement à la sémiosis ou interpréter un texte.[4] »

L'approche du document en termes d'interprétant favorise, semble-t-il, *un mode utilitaire* des sources ; ceci semble cohérent avec notre autre conjecture, où la source se définit en fonction de ses possibilités de détournement. Cependant, et notamment parce que l'indexation doit maintenir un lien entre les différents espaces de représentations (des auteurs, des indexeurs, des utilisateurs), il faut pouvoir stopper « l'entraînement de la sémiosis » ; autrement dit, il faut pouvoir circonscrire l'univers de discours[5], c'est-à-dire l'espace des utilisations qu'il est nécessaire de proposer aux utilisateurs : c'est en ce sens que nous proposerons la notion de discours documentaire* (*infra*, chapitre IV).

Mais avant, nous devons préciser le rapport qu'entretiennent utilisation et interprétation (B1). De ce point de vue s'éclaire la problématique du traitement documentaire de l'image (B2).

B1 - Quelles sont les relations qu'entretiennent utilisation et interprétation d'un texte ? Sont-elles exclusives l'une de l'autre ?

Sur ce point, Umberto Eco note que, si « leur libre utilisation n'a rien à voir avec leur interprétation [...] toute lecture est toujours un mélange des deux », dans la

---

[1] Eco 1985 [1979] et Eco 1992 [1990].
[2] Eco 1992 [1990], p. 39.
[3] On peut tout à fait considérer, à la suite de Bourdieu [1982] par exemple, que seul l'usage fait le sens et donc que l'interprétation d'un texte n'est rien d'autre que l'utilisation qui en est faite. Cette position revient, nous semble-t-il, à nier le rôle de la langue dans les faits d'interprétation. C'est pourquoi nous ne la retiendrons pas dans cette recherche.
[4] Eco 1985 [1979], p. 74. (C'est nous qui soulignons).
[5] *Ibid.*, p. 45 : la limite logique à l'entraînement sémiotique est l'univers de discours, « un univers de faits limité ».

mesure où est toujours présupposée « une référence au texte-source, du moins en tant que prétexte[1] ».

C'est alors le type de lecture qui est fait de la source qui détermine le rôle que joue le document en tant qu'interprétant : mode d'interprétation ou mode d'utilisation d'une source ?

C'est ici à la distinction établie par Escarpit[2] entre lecture objective et lecture projective que nous nous référons :

- dans la lecture objective, on cherche à « épuiser l'entropie du texte, c'est-à-dire à énoncer toute l'information qu'il contient, à le rendre entièrement connu, de sorte que tout nouveau balayage ne produira que des événements prévisibles[3] ». En reprenant la problématique d'Eco, on peut dire que ce type de lecture fige la source dans l'une de ses utilisations ;

- dans la lecture projective, il s'agit d'effectuer une « mise en mémoire[4] » d'une source dans un document de façon à ce qu'elle reste « disponible pour une lecture c'est-à-dire pour une exploration libre de toute contrainte événementielle ou chronologique[5] ». Cette lecture consiste à réaliser une double opération : « sélectionner » et « associer[6] ». Par la sélection, cette lecture introduit le facteur « oubli », que nous traduisons comme étant l'oubli des contextes de production d'une source. Par l'association, cette lecture introduit des « possibilités d'évocation analogique[7] ». Dans les termes d'Eco, on peut dire que cette lecture fournit la possibilité d'utilisation diverse mais réglée d'un texte (contraintes liées à celles de l'analogie) .

Dans les deux types de lecture, il y a, avant toute décision d'utilisation, nécessairement une interprétation de la source ; cette interprétation pèse sur les utilisations possibles. En ce sens, l'utilisation est toujours seconde et toujours déterminée par l'interprétation, qui, elle, est toujours première, sauf dans le cas particulier des objets non textuels où l'utilisation ne nécessite pas une « interprétation » proprement dite. C'est, nous semble-t-il, pour cette raison que les objets non textuels ne deviennent documents, candidats à l'indexation, qu'une fois affectés d'un texte qui, lui, peut faire l'objet d'une interprétation. Nous abordons succinctement cette problématique ci-dessous.

B2 - La problématique du rapport utilisation/interprétation se trouve au cœur du traitement documentaire de l'image[8]. En effet, l'utilisation de l'image peut complètement supplanter l'interprétation[9] : se posent, alors, de façon cruciale, le

---

[1] Eco 1992 [1990], p. 46-47.
[2] Escarpit 1991, p. 129.
[3] *Id.*
[4] *Ibid.*, p. 150.
[5] *Ibid.*, p. 63.
[6] *Ibid.*, p. 152.
[7] Par l'analogie se crée la « production d'un nouvel événement sous l'effet d'un nouveau stimulus qui, d'une part, suscite une réponse libre, d'autre part, peut réactiver par analogie et non plus par identité la trace d'une expérience passée », Escarpit 1991, p. 151.
[8] Pris ici comme exemple d'objet non textuel soumis à l'indexation. Il faudrait analyser, dans le détail, le comportement des autres objets non textuels du point de vue du rapport utilisation/interprétation.
[9] Conférence Hudrisier ENSSIB (Villeurbanne), mars 1996.

problème de l'absence de contraintes interprétatives et le danger du détournement incontrôlé des images. On connaît l'exemple fameux de l'utilisation intempestive d'une des photographies de Robert Doisneau[1] : la scène photographiée est celle d'une discussion, dans un café de la rue de Buci à Paris, entre une jeune étudiante et son professeur d'architecture. Utilisée une première fois comme support d'illustration pour une campagne anti-alcoolique, la même image apparaît, peu de temps après, dans un magazine « à scandale » avec pour légende : « Prostitution aux Champs-Élysées ». Reprenant ce même exemple, Odile Le Guern souligne que seule la formulation linguistique (légendes et/ou formules d'indexation) peut neutraliser l'interprétation et contraindre ainsi les utilisations d'une image[2]. Autrement dit, l'image seule, décontextualisée, ne supporte aucune contrainte d'utilisation ; pourvue d'un texte, l'image est contextualisée et c'est cette contextualisation (par le texte) qui établit des contraintes d'utilisation.

La problématique de l'indexation des images mise en lumière par Odile Le Guern montre que, lorsqu'il n'y a pas de texte, interprétation et utilisation ne peuvent être distinguées : il n'y a pas alors d'indexation possible ; ce qui dessine en creux le processus de l'indexation : il repose sur la distinction des deux modes d'interprétation et d'utilisation. Pour proposer des utilisations des sources (soit unique dans le cas de la lecture objective, soit multiple dans celui de la lecture projective), l'indexation doit exploiter préalablement la nature textuelle, ou plutôt discursive, des documents (niveau de l'interprétation).

L'emprunt, au modèle de Peirce, de la notion d'interprétant nous permet de dégager une problématique peu abordée dans le discours classique : l'indexation vise-t-elle, via son référent, à fournir des modes interprétation ou des modes d'utilisation des sources ?

Nous avons fait l'hypothèse que, si l'indexation devait plutôt fournir des utilisations des sources, elle devait se donner le moyen de contraindre ces utilisations : en ce sens, nous avons fait l'hypothèse que l'indexation créait un niveau de discours particulier – le discours documentaire – dans lequel les utilisations des documents pouvaient être contrôlées. Dans ce cadre, quelles sont les propriétés dont doit être pourvu le document ?

C - Propriétés du document

Nous ne faisons ici qu'amorcer une réflexion sur les propriétés dont doit être pourvu le document pour « fonctionner comme » l'interprétant d'une source. Nous en proposerons des prolongements dans le chapitre IV.

(i) Propriété de recontextualisation

Si le document est construit sur la base de la décontextualisation d'une source, il doit être recontextualisé, sous peine de rester au stade du semi-document : il s'agit là d'un « effet de compensation » dans les termes d'Escarpit[3]. La recontextualisation du document doit se faire de façon à permettre de créer des

---

[1] Voir Hudrisier 1984 et O. Le Guern 1989.
[2] O. Le Guern 1989, p. 427.
[3] Escarpit 1991, p. 125 : « On notera que le temps, dont l'effet est compensé lors de la constitution du document, doit être réintroduit sous forme de *mouvement* pour que l'information soit restituée au destinataire ».

« évocations analogiques » : un document n'est tel que s'il permet d'établir des liens avec d'autres documents sous une forme qui ne soit pas celle de la répétition à l'identique. C'est pourquoi tous les organismes documentaires ne construisent pas nécessairement, à partir d'une même « réalité éditoriale », les mêmes documents[1] : tout dépend des autres documents déjà présents. Ces documents constituent un « contexte » les uns pour les autres ; sur ce point, la recontextualisation d'une source qui lui permet de fonctionner comme un document repose sur une mesure de « compatibilité[2] » entre la source et les documents déjà sélectionnés.

(ii) Propriété de stabilité

En mettant en regard document et événement, Escarpit souligne l'effet de stabilisation que peut donner un document. Si le document a pour propriété celle de cette stabilité, il ne l'a que dans le cadre de l'espace documentaire[3], qui stabilisant la sémiosis, stabilise du même coup un état du monde, qui peut alors se donner comme « réalité commune » aux indexeurs et aux utilisateurs. C'est parce que la propriété de stabilité n'est valable que dans le cadre de l'espace, du discours documentaire, qu'elle nous semble être une propriété distinctive du document.

Nous avons, dans un premier temps (§ III.1.1 et § III.1.2 ci-dessus), tenté de distinguer les différents processus en jeu dans la construction du document en indexation, en adoptant deux points de vue différents :

- *le point de vue de la source, abordé sous l'angle de la théorie de l'énonciation, laisse apparaître la notion de contexte (contexte dans lequel une source est produite) comme ce qui distingue une source d'un document ;*

- *le point de vue du document, appréhendé dans le cadre du modèle de Peirce, a permis, lui, de préciser ce qui rapproche source et document : ils se trouvent dans un double rapport d'interprétation et d'utilisation.*

Nous proposons désormais de synthétiser, sous la forme d'un schéma, le processus de la création d'un document à partir d'une source. Par cette schématisation, nous montrerons que la construction du document telle que nous l'entendons fait partie intégrante du processus de l'indexation.

III.1.3 - LA CONSTRUCTION DU DOCUMENT : UNE OPÉRATION À DOUBLE DÉTENTE

Nous commençons par exposer notre schématisation de la construction du document en indexation (A), que nous commentons ensuite sous trois aspects :

(B) commentaire sur la notion de contexte : la schématisation proposée repose sur la notion de contexte, notion floue s'il en est, que nous essayerons néanmoins de préciser pour ce qui concerne son rôle en indexation ;

---

[1] Voir précédemment, dans ce chapitre § 1.3.1., le résultat de notre expérimentation.
[2] Ce point est développé dans le chapitre IV.
[3] Espace artefactuel de non-changement dans les termes d'Escarpit [1991, p. 125] : « Un objet stable reste emporté par le temps, la stabilité n'étant que le non-changement des relations qui définissent sa configuration caractéristique (son *pattern*) aux yeux d'un observateur emporté par un temps qui est, par convention pratique, supposé le même ».

(C) commentaire sur la notion de processus : cette schématisation suppose une compréhension de la notion de processus en indexation au sens « fort » ; il s'opère, en indexation, notamment par la création du document, une transformation d'objets ;

(D) commentaire sur la visibilité de ce processus : cette schématisation rend visible un aspect de l'indexation (la mise en document) qui ne bénéficie généralement que d'une très faible visibilité : nous aborderons pour finir cet aspect.

A - Une opération à double détente

Si la mise en document opère une « recontextualisation » dans un cadre particulier à préciser, une opération, symétrique et antérieure, de « décontextualisation », doit être postulée. La phase de décontextualisation, qui correspond à la notion classique de « sélection documentaire », est donc indissociablement liée aux conditions d'existence du document. Dès lors, notre thèse consiste à appréhender l'indexation comme un processus à double détente, composé de deux opérations symétriques et inverses de contextualisation[1].

Pour fixer les idées, on pose, en première approximation, la schématisation de notre thèse sous la forme suivante :

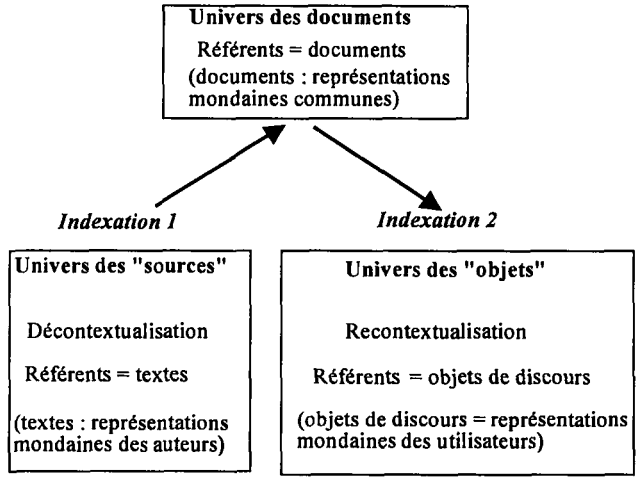


Figure 3 – Mécanisme « à double détente » de l'indexation

Dans ce schéma, l'indexation crée son référent (le document) en procédant à une double opération de contextualisation :

[1] On est assez proche en cela de la position défendue par Escarpit qui parle, lui, de déconstruction et de reconstruction du texte, Escarpit 1991, p. 166 : l'analyse documentaire « doit procéder à une réécriture du texte qui n'est pas simplement linéaire ou quantitative, mais qui comporte une déconstruction, puis une reconstruction systématique du texte ».

- la première est une décontextualisation : la source est détournée de son utilisation initiale ;

- la seconde est une recontextualisation : la source prend de nouvelles potentialités interprétatives, intégrée dans un ensemble que l'on nommera « discours documentaire ». Dans ce cadre se dispose l'ensemble des usages possibles d'un document (des détournements autorisés d'une source).

L'enjeu de l'indexation est donc de réaliser un changement de contexte : cette notion, qui met en valeur le processus (au sens fort) de l'indexation, mérite d'être précisée.

B - Notion de contexte en indexation

La notion de contexte est, en linguistique, mal définie. Elle est, d'après Kerbrat-Orecchioni[1], à la fois « problématique[2] » et « indispensable[3] ». Rastier, qui se livre à une semblable remarque[4], insiste, lui, sur la nécessité d'approcher la notion de contexte en termes de palier : syntagme, période, section, texte constituent chacun un « palier de contextualité ».

Rastier traite ce faisant la notion de contexte linguistique[5] (dit aussi co-texte) que l'on oppose généralement à la notion de contexte extra-linguistique[6] (dit aussi contexte situationnel)[7]. Kerbrat-Orecchioni s'attache, elle, plus à ce deuxième type de contexte.

Ce que ces deux auteurs mettent en lumière, c'est que le contexte n'est pas un donné, mais un construit. À ce titre, on peut dire qu'un contexte, c'est « l'ensemble des représentations que les interlocuteurs ont du contexte[8] ». En ce sens, un contexte est toujours « choisi », pour reprendre les termes de Rastier[9], et ce choix constitue un « acte décisif » : « Alors que le texte appartient au "donné" empirique, tel qu'on choisit de le décrire, le contexte est *choisi*, donc soumis comme tel à des conditions herméneutiques. Dans l'interprétation des mots, comme dans

---

[1] Kerbrat-Orecchioni 1996, p. 39-60.
[2] Latraverse cité *in* Kerbrat-Orecchioni 1996, p. 39 : « La notion de contexte est d'une telle souplesse et d'un accueil si généreux qu'il est difficile de considérer qu'elle a des frontières suffisamment établies pour jouer un rôle théorique non équivoque ».
[3] *Id.* : « Il n'est guère de théorie ou d'approche sémantique des langues naturelles qui en fasse l'économie ».
[4] Rastier 1994, p. 64 : « On ne dispose pas en linguistique de théorie générale du contexte. Si la notion de contexte est souvent évoquée, elle est rarement définie ».
[5] Ce type de contexte peut être entendu comme l'« ensemble du texte qui entoure un élément de la langue (mot, phrase, fragment d'énoncé) et dont dépend son sens, sa valeur », Kerbrat-Orecchioni 1996, p. 40.
[6] Ce type de contexte peut être entendu comme l'« ensemble des circonstances dans lesquelles s'insère un fait ». *Id.*
[7] Le problème de cette distinction, même si elle est importante à maintenir sur un plan méthodologique, est, comme le remarque Kerbrat-Orecchioni, que les deux types de contexte sont en fait dans un rapport de « vases communicants », voir Kerbrat-Orecchioni 1996, p. 40.
[8] *Ibid.*, p. 41.
[9] De même pour Kerbrat-Orecchioni, la dimension du contexte à prendre en compte « dépend évidemment des cas ». *Ibid.*, p. 40.

l'interprétation des textes, le choix du contexte est un *acte décisif, qui doit être déterminé par une stratégie.* (C'est nous qui soulignons).[1] »

Au risque de remplacer une notion floue par une autre notion floue, nous essayerons d'approcher la notion de contexte en indexation sous l'angle de la notion de choix stratégique.

Les questions qui se posent à une approche du contexte sont alors formulées dans d'autres termes :

(i) La question de la dimension du contexte à prendre en compte

Le contexte d'une source peut être approché par la notion de « contexte situationnel », contexte extra-linguistique des « circonstances ». Le contexte du document relève, lui, d'abord du contexte linguistique, ensuite du contexte situationnel.

Le contexte du texte d'un document est en effet constitué des descripteurs affectés à ce document : on peut parler à ce titre de modification du contexte linguistique du document. Ces descripteurs fonctionnent comme des relais lexicaux entre plusieurs documents ; ils contribuent alors à créer un nouveau contexte d'interprétation, un nouveau contexte situationnel : le contexte d'un document est ainsi constitué des autres documents qui partagent les mêmes descripteurs. C'est sur ce point qu'apparaît nous semble-t-il la notion de stratégie documentaire : c'est elle qui détermine le choix du regroupement des documents comme celui des textes dans un même document ; ce choix s'exprime sous une forme lexicale, par les descripteurs, mais n'est que le résultat d'une stratégie qui ne se joue pas au seul niveau des mots.

En ce sens, on dira que l'indexation opère le passage d'un contexte situationnel à un autre, par modification du contexte linguistique des documents (les descripteurs), et qu'elle obéit ce faisant à une stratégie que nous aurons à préciser.

(ii) La question du rapport entre un texte et un contexte

Kerbrat-Orecchioni met en valeur la relation dialectique qui caractérise le rapport entre texte et contexte : « le discours façonne le contexte autant que le contexte façonne le discours[2] », et ce, parce que texte et contexte peuvent être de même nature, précise Kerbrat-Orecchioni : « une unité donnée n'est pas en soi élément du texte ou du contexte[3] ».

Là encore, pour ce qui est des faits d'indexation, il nous semble important de relever que, si une collection documentaire peut se comprendre en termes de situation d'interprétation (de contexte situationnel) pour des utilisateurs, c'est parce que la stratégie documentaire permet aux documents qui s'y trouvent de se transformer et/ou de s'éclairer les uns les autres de façon « légitime », pour reprendre les termes d'Eco. Cet éclairage qui permet à un texte de se donner comme le contexte d'un autre texte repose, selon Kerbrat-Orecchioni, sur la notion

[1] Rastier 1994, p. 65.
[2] Kerbrat-Orecchioni 1996, p. 49.
[3] *Ibid.*, p. 43.

d'« événement focal », mise en valeur par Duranti & Goodwin[1]. Cette notion d'« événement focal » qui permet, par le biais d'un texte T_1, de focaliser un aspect d'un texte T_2 (T_1 joue alors un rôle de contexte pour T_2) nous semble être à l'œuvre en indexation : nous y reviendrons dans le chapitre V.

(iii) La question de la production et de l'interprétation du contexte

Cette question nous semble pouvoir être également envisagée sous l'angle du choix stratégique.

Du côté des indexeurs, il y a production d'un nouveau contexte situationnel (univers des documents dans le schéma), production déterminée par une stratégie (que nous essayerons de définir) ; du côté des utilisateurs, il y interprétation du contexte situationnel proposé par les documents, interprétation elle aussi déterminée par l'adoption d'une stratégie (ici une stratégie de recherche, qui ne relève pas de notre étude). Il importe peu que stratégie de production et stratégie d'interprétation des contextes soient ou pas de même nature ; il importe surtout, nous semble-t-il, de constituer les actes des différents acteurs de l'indexation sous forme de stratégie. En effet, comme le note Eco, « la coopération textuelle est un phénomène qui se réalise, nous le répétons, entre deux stratégies discursives et non pas entre deux sujets individuels[2] ». Ce qu'Eco appelle ici « coopération textuelle », c'est l'acte de lecture. La stratégie de l'auteur est en indexation (dans l'univers des documents) mise au second plan, le premier plan étant occupé par la stratégie des indexeurs : c'est elle qu'il importe, selon nous, de dégager. L'indexeur se trouve pris alors dans le mouvement de la coopération textuelle, de façon plus décisive que ne le laisse supposer l'image de l'indexeur-médiateur. S'il est en effet souvent question, dans le discours classique, de la stratégie (ou des stratégies) des utilisateurs, il est peu question de la stratégie (ou des stratégies) des indexeurs.

Nous avons proposé d'approcher le contexte en indexation comme élément d'une stratégie plus globale, menée par les indexeurs pour donner à lire et à utiliser des sources dans des sens et des usages qui puissent être « nouveaux ». C'est cette stratégie qui permet de considérer le processus de l'indexation dans le sens fort de « transformation ».

C - Notion de processus

Aussi triviale qu'elle puisse paraître, l'intégration d'une source dans une collection documentaire ne s'effectue ni de façon évidente ni de façon neutre : par l'utilisation des deux modèles de l'énonciation et de l'interprétant, nous avons cherché à montrer que cette intégration supposait une réelle transformation des objets. En effet, la modification du contexte situationnel qui signe le passage de la source au document introduit les éléments d'une lecture qui seuls permettent à une source d'être un document, c'est-à-dire de pouvoir être interprétée et utilisée à d'autres fins que celle qui lui avait été initialement attribuée. C'est en ce sens que l'indexation constitue un processus : nous nous attacherons désormais dans cette recherche à le préciser.

---

[1] « The context is thus a frame that surrounds the event being examined and provides resources for its appropriate interpretation », repris de Kerbrat-Orecchioni 1996, p. 43-44.
[2] Eco 1985 [1979], p. 78.

La difficulté tient à ce que ce processus n'est pas rendu visible : le remplacement d'un contexte situationnel par un autre ne se voit pas ; la phase de décontextualisation de la source reste peu visible, puisqu'au bout du compte, il y a toujours un contexte situationnel.

D - Défaut de visibilité

Le marquage le plus net de ce que l'on nomme ici la recontextualisation est l'« indexation », comprise dans le sens étroit d'attribution, à un document, de descripteurs : l'indexation ainsi entendue *matérialise*, en effet, sur un plan linguistique, les opérations effectuées en amont sur les sources, autrement dit la sélection documentaire.

Si l'on tente de se dégager des « effets » pour capter les « fondements », on peut soutenir l'hypothèse que la sélection documentaire (ou, plus exactement, la sélection des sources) relève du processus de l'indexation, même si la matérialisation y est reportée. On rejoint, ce faisant, les conclusions de Dubois et Mondada, émises suite à une expérience de tri de photographies : « Pour résumer ce point, on dira que, même à un niveau non explicitement verbal (le tri ne demandait aucune lexicalisation), la discrétisation du monde en catégories n'est absolument pas donnée *a priori*, mais varie selon les activités cognitives des sujets qui les opèrent.[1] »

De même que dans le compte rendu de cette expérience, on peut dire que, dans l'indexation, la discrétisation du monde en objets (la création des documents), qui n'est le plus souvent pas verbalisée, constitue une opération qui influe directement sur la décision *in fine* du choix de tel ou tel descripteur.

Les deux phases du processus de l'indexation n'offrent donc pas la même visibilité. En effet, seule la seconde phase (dite de recontextualisation) se réalise sous une forme linguistique, mais cette matérialisation ne correspond en fait qu'à l'établissement préalable et non verbalisé de liens posés entre les documents. Le fait que, dans le discours classique, la sélection documentaire ne soit pas si évidemment comprise dans le mécanisme de l'indexation tient sans doute à ce que la phase de recontextualisation, en tant que phase de désignation, soit la seule à être dotée d'une verbalisation. À ne considérer que la face visible de l'indexation, on court cependant le risque de mésinterpréter le rôle des descripteurs en indexation : matérialisations d'une analyse documentaire réalisée en amont sur la base de principes de regroupement textuel, ils ne peuvent, à proprement parler, rendre compte du contenu d'une source ; ils peuvent, en revanche, donner à lire, à interpréter et à utiliser cette source.

Au début du paragraphe III, notre objectif était de préciser comment l'indexation pouvait se comprendre comme un acte de référenciation, c'est-à-dire comme un acte établissant un double processus de discrétisation et de stabilisation.

Nous avons essayé de dégager, en III.1, le processus de discrétisation que l'indexation opère pour construire son « monde », celui des documents : nous avons établi que ce processus de discrétisation était réalisé en indexation sur la base d'une transformation de contextes, transformation qui donne la possibilité d'utilisations nouvelles.

---

[1] Dubois et Mondada 1995, p. 288.

Nous nous intéressons, dans le paragraphe suivant, au second aspect de l'acte de référenciation : le processus de stabilisation. Nous montrerons alors comment l'indexation peut construire des effets de stabilité référentielle.

III.2 - Construction de l'effet de stabilité référentielle en indexation

Pour appréhender l'indexation sous l'angle d'un processus de stabilisation, nous ne nous intéresserons pas, comme précédemment, à l'acte de référence lui-même mais aux propriétés référentielles des expressions linguistiques sollicitées dans la réalisation de cet acte.

Nous avons vu précédemment[1] que, s'il y avait stabilité référentielle par le biais d'une stabilité linguistique (cas où un mot renvoie à un objet et un seul), il ne s'agit là que d'un *effet*, produit de façon typique par l'interprétation d'un type d'unité linguistique particulier : le nom propre[2]. C'est sur ce type d'unité que nous centrons nos propos dans cette partie de l'étude et ce, sous deux angles :

- nous chercherons d'une part à expliquer ce qui permet à un nom propre de créer un effet de stabilité référentielle : pour cela, nous empruntons au logicien Kripke[3] la notion de « rigidité » telle qu'il l'a établie[4] ;

- nous montrerons d'autre part que la pratique d'indexation, qui recourt largement à l'utilisation de noms propres, n'est pas sur ce point une pratique isolée. D'autres pratiques révèlent un recours massif à ce type d'unité : il y a là, nous semble-t-il, un élément susceptible de contribuer aux recherches portant sur l'analyse des effets stabilisateurs des pratiques[5].

On remarquera par ailleurs que l'effet de stabilisation que peut obtenir l'indexation, *via* l'utilisation de propriétés linguistiques particulières, ne rend que plus délicate la mise en faillite du modèle réaliste en indexation ; en effet, l'obtention de cet effet stabilisateur permet au discours classique de pouvoir, ne serait-ce que partiellement, réaliser son pari : faire correspondre aux mêmes mots toujours les mêmes choses. Reste que le discours réaliste ne peut expliquer pourquoi, dans tel cas, son pari est tenu et pourquoi, dans tel autre cas (quand les descripteurs sont des noms communs), il l'est moins : c'est précisément ce que nous permet de faire le modèle non réaliste de la référence que nous avons adopté.

---

[1] Voir § I.3.2.

[2] Tout comme Gary-Prieur [1994, p. 7], on entendra ici, par noms propres, les « noms propres prototypiques », de type nom de personnes, nom de lieux, nom de marques, etc., laissant de côté les cas problématiques (comme les noms de peuples, écrits avec une majuscule). Par ailleurs, nous nous attacherons aux seuls cas où le nom propre est employé sans déterminant, ignorant les emplois du type : « Je croyais que Ramiz Alia pouvait devenir *le* Gorbatchev albanais » (repris de Gary-Prieur 1994, p. 38).

[3] Kripke 1982 [1972].

[4] Nous n'abordons pas ici les approches strictement linguistiques du nom propre, bien qu'elles se réfèrent toutes, plus ou moins explicitement, aux positions de Kripke. Il y a débat entre ceux pour qui le nom propre peut faire l'objet d'une description sémantique (analyse en termes de présupposition) et ceux pour qui elle ne le peut pas (le nom propre n'a pas de signification lexicale) : pour une revue de la question, on peut se reporter par exemple à Kleiber [1981] et Gary-Prieur [1994].

[5] Dubois et Mondada [1995, p. 292-297] proposent d'engager une réflexion sur les moyens, cognitifs et linguistiques, par lesquels un effet de stabilisation du monde peut être obtenu.

II.2.1 - PROBLÉMATIQUE DE LA « RIGIDITÉ »

De nature logique et philosophique, et non linguistique[1], la réflexion de Kripke s'inscrit dans le cadre d'un vaste débat, dont nous ne rappelons ici que les principaux enjeux, en reprenant : la question initiale de ce débat et la solution proposée par Kripke (notion de rigidité) en A ; les corollaires au principe de la rigidité : la notion de « mondes possibles » (B), celle de désignateur rigide (C) et la dimension discursive du nom propre (D).

A - Question initiale : les propriétés référentielles du nom propre

Kripke s'oppose à Searle qui, s'inspirant de Wittgenstein, pose que les noms propres, comme les descriptions définies[2], ont un « sens », qui connote, pour les descriptions définies, une propriété bien déterminée et, pour les noms propres, un faisceau mouvant de propriétés : cette seule différence, de degré, entre descriptions définies et noms propres est également retenue par Frege et Russel qui traitent alors les noms propres comme des « descriptions déguisées ». Dans ce cadre, le référent d'un nom propre « est l'objet qui possède un nombre suffisant de propriétés qu'on associe à ce nom dans la communauté linguistique[3] ».

Kripke reprend la position de Mill pour soutenir une thèse inverse : il pose, comme lui, que les noms propres se distinguent des descriptions définies, en ce qu'ils désignent un objet *indépendamment des propriétés que cet objet possède*. Mais, alors que Mill utilise la notion de connotation pour définir, par la négative, le nom propre, Kripke conçoit le principe de « rigidité » pour définir de façon positive le nom propre et rendre compte de l'effet de stabilité référentielle propre à certaines descriptions définies.

C'est dans ce cadre que Kripke propose la notion de « désignateur rigide » qu'il oppose à celle de « désignateur accidentel ».

Par « désignateur », Kripke entend « un terme commun couvrant à la fois les noms et les descriptions[4] », étant entendu que le nom est, dans ses termes, « un nom propre, c'est-à-dire le nom d'une personne, d'une ville, d'un pays, etc.[5] » Dans ce cadre, Kripke distingue donc :

- les désignateurs rigides : « expression dont la dénotation* ne varie pas, quel que soit le monde considéré[6] » ;

- les désignateurs accidentels : « expression dont la dénotation varie selon le monde de référence, qui ne désigne pas le même objet dans tous les mondes possibles[1] ».

---

[1] Il n'empêche que la notion de « désignateur rigide » a été largement reprise en linguistique, voir Gary-Prieur 1994, p. 14-25.
[2] Depuis Russel, les logiciens entendent par « descriptions définies » « les expressions comportant un nominal (nom, nom+adjectif, nom+relative, nom+complément) accompagné d'un article défini » ; des expressions comme « le livre » ou « le livre que j'ai prêté » sont analysées en logique comme des descriptions définies ; repris de Ducrot et Schaeffer 1995, p. 306.
[3] Récanati 1983, p. 107.
[4] Kripke 1982 [1972], p. 13.
[5] *Id.*
[6] Récanati 1983, p. 109.

B - Notion de « monde possible »

Le principe de rigidité est étroitement associé à la notion de « mondes possibles[2] » qui relève, elle, de la logique modale[3].

Soit les deux exemples suivants[4] contenant chacun l'opérateur modal « aurait pu » :

(1) Le président de la République aurait pu être un homme de gauche.

(2) Chirac aurait pu être un homme de gauche.

En (1), l'énoncé est ambigu : il réfère soit au président actuel (référence au monde réel[5]) que l'on projette dans un « monde possible » où il aurait été de gauche, soit au président d'un monde possible (référence à un monde possible) qui, lui, serait de gauche[6]. Selon le monde de référence considéré (monde réel ou monde possible), le référent varie en (1) : c'est pourquoi la description définie « le président de la République » est dite « désignateur accidentel ». En (2), il n'y a pas d'ambiguïté : que l'on parle du monde réel ou d'un monde possible, le référent est toujours Chirac[7] ; en cela, le nom propre est un « désignateur rigide ».

Comme le signale Kripke, la rigidité du nom propre correspond à un primat du monde réel : c'est toujours le référent du monde réel qui est désigné même si est imaginée une situation contrefactuelle où le référent du nom propre ne présente aucune de ses propriétés.

Soit, par exemple[8], l'énoncé (3) « Imaginons que Hitler ait été doux comme un agneau » ; c'est précisément parce que Hitler garde ses propriétés du monde réel, à savoir être responsable d'un génocide, que l'énoncé est interprétable[9].

---

[1] Récanati 1983, p. 109..

[2] La question de savoir si la notion de « monde possible » suggère nécessairement un engagement ontologique est ouverte ; Kripke n'en dit rien de précis. Certains linguistes commentateurs de Kripke proposent d'analyser la notion de « monde possible » en termes de discours : ainsi, selon Gary-Prieur, les mondes possibles « n'ont pas d'existence extérieure au discours qui les met en place », Gary-Prieur 1994, p. 21. C'est cette lecture que nous privilégions dans cette recherche, voir aussi le chapitre IV.

[3] Il est hors du cadre de cette étude de préciser ce qu'on peut entendre par « logique modale » ; nous nous efforcerons simplement, dans le paragraphe qui suit, d'exposer les éléments pertinents à la compréhension de notre démarche.

[4] Les exemples présentés ci-dessous ainsi que leurs commentaires sont repris de Récanati 1983, p. 108-109 ; les exemples sont, pour des raisons de commodité de présentation, mis au goût de l'actualité politique actuelle.

[5] Ici encore, on peut considérer que la notion de « monde réel » est chez Kripke de nature discursive, voir Gary-Prieur 1994, p. 21 : « Même si ce n'est pas dit explicitement sous cette forme dans le texte de Kripke, je crois que c'est, en dernière analyse, l'acte d'énonciation qui fonde le " monde réel " auquel s'opposent les mondes possibles ».

[6]. Plus précisément, la première lecture porte sur l'objet lui-même (« le président de la république »), c'est une lecture *de re* ; la seconde lecture porte sur l'ensemble de l'énoncé, elle est dite *de dicto*. Nef 1991, p. 102.

[7] Le nom propre neutralise l'ambiguïté des lectures *de re* et *de dicto*.

[8] Repris de Récanati 1983, p. 109.

[9] *Id.* On voit la différence avec l'exemple suivant : « Imaginons que le chef de l'Allemagne nazie ait été doux comme un agneau ».

Par conséquent, « Dire que la relation de désignation qui associe le nom propre à l'objet est "rigide", c'est dire qu'elle n'est pas affectée par le changement du monde de référence, et dire cela, c'est dire qu'elle n'est pas fonction des propriétés de l'objet, dans la mesure où ce sont ces propriétés qui varient d'un monde à l'autre.[1] »

Cette permanence de la référence maintenue par le nom propre à travers tous les univers de référence imaginables n'est pas sans nous rappeler l'ambition de l'indexation qui cherche à maintenir, par une même forme linguistique, une même stabilité référentielle[2].

C - Notion de désignateur rigide

Ce principe de rigidité, illustré de façon exemplaire par le nom propre, est également à l'œuvre dans le fonctionnement référentiel d'autres types d'expressions linguistiques. En effet, dans la théorie de Kripke, certaines descriptions définies peuvent également fonctionner comme des désignateurs rigides (c'est d'ailleurs sur ce point que Kripke se sépare de Mill).

Par ailleurs, il faut également préciser que, si la démonstration de la rigidité est particulièrement visible appliquée aux contextes modaux, Kripke ne la soutient pas moins dans tous les autres types de contexte[3]. L'exemple suivant illustrera la position de Kripke sur ces deux derniers aspects (description définie comme désignateur rigide et absence d'opérateur modal) :

(4) *π est censé être le rapport de la circonférence d'un cercle à son diamètre*, où « π » et « le rapport de la circonférence d'un cercle à son diamètre » sont des désignateurs rigides même s'ils ne le sont pas au même titre[4].

En effet, Kripke[5] distingue la rigidité *de jure* (par stipulation), caractéristique des noms propres, et la rigidité *de facto*, qui couvre le cas où les descriptions définies ne sont vraies que d'un seul objet, quel que soit l'univers envisagé. Kripke pense surtout aux descriptions mathématiques : ainsi, le référent de la description définie « le plus petit nombre premier » est le nombre 2, dans tous les mondes possibles. La rigidité vient donc dans ce cas de la nature du référent désigné. La rigidité étant dans ce cas *de facto*, l'identification du référent se fait hors du contexte de l'énoncé, alors que, pour les désignateurs rigides *de jure*, le référent est directement identifiable dans l'énoncé. Nous revenons sur la portée de cette distinction dans le chapitre V.

---

[1] Récanati 1983, p. 110.

[2] Notons, cependant, comme le fait remarquer Kripke, que la notion de rigidité ne suppose pas de critère d'identité à travers les mondes ; l'identité référentielle est une *conséquence* de la rigidité, voir l'enjeu de la distinction étudiée dans le chapitre IV.

[3] Kripke 1982 [1972], p. 162-166.
Hors de contextes modaux, les deux exemples (1) et (2) s'opposent tout autant, mais l'opposition s'exprime ici en termes de valeur de vérité et non en termes de potentialité : « Le président de la République est un homme de gauche » (si c'est Chirac, c'est faux ; si c'est Mitterrand, c'est vrai) et « Chirac est un homme de gauche » (c'est faux, quel que soit le monde de référence).

[4] Kripke 1982 [1972], p. 48-49.

[5] *Ibid.*, p. 173.

D - Dimension discursive du nom propre

La théorie de la rigidité énoncée par Kripke montre que la langue peut envisager des cas de « référence directe [...] qui court-circuite la machine intensionnelle[1] », c'est-à-dire des cas où les unités linguistiques établissent une relation référentielle stable sans que l'on ait à se préoccuper de leur « sens ». Les désignateurs rigides illustreraient donc un cas d'autonomie de la relation référentielle, ce qui ne signifie pas que cette autonomie existe en soi ; elle est, en effet, issue non du réel mais de situations discursives[2]. Selon Kripke, la fixation des référents des désignateurs rigides *de jure* se fait en effet par le biais d'une « chaîne causale » : « La référence semble finalement déterminée par le fait que le locuteur fait partie d'une communauté de locuteurs qui utilisent le nom. Le nom lui a été transmis grâce à une tradition, de maillon en maillon.[3] »

Après un « baptême initial » au cours duquel une description définie est associée à un individu et à un nom propre, se met en place une « chaîne causale » qui correspond à la transmission du nom propre d'un locuteur à l'autre. Mais, comme le précise Kripke, « la manière dont la référence d'un terme est fixée n'a rien à voir avec sa signification[4] ».

C'est cette particularité de la fixation de la référence, réalisée une fois pour toutes après un baptême initial, qui distingue les pronoms des noms propres : si, comme le note Milner, la référence virtuelle des deux types d'unité « contient déjà la mention de l'unité en cause[5] », cette référence virtuelle est circulaire dans le cas du pronom, elle ne l'est pas dans le cas du nom propre ; autrement dit, la référence du nom propre n'est pas suspendue à l'énonciation, elle existe en dehors d'elle précisément en vertu de la chaîne causale qui constitue sa référence : « [Les noms propres] donnent apparemment lieu à la circularité : la définition de *Jean*, c'est d'être dit *Jean* mais en fait c'est une fois pour toutes et sans qu'il faille répéter l'énonciation qu'un sujet reçoit un nom propre. Ainsi la classe des êtres appelés *Jean* est objectivement déterminable, sans intervention à chaque énoncé singulier d'un sujet énonciateur.[6] »

Cette propriété de non-circularité dégagée par Milner appuie la position de Kripke précédemment rapportée sur la distinction entre fixation référentielle et signification du nom propre.

Le principe de « rigidité » établi par Kripke nous paraît éclairer avantageusement le mécanisme de l'indexation. En effet, la notion de « désignateur rigide » incarne la possibilité d'établir dans la langue une permanence référentielle transcendant la diversité des mondes de référence, et cela grâce à l'insensibilité que témoigne le nom propre à l'égard des propriétés de l'objet qu'il désigne. Une telle possibilité

---

[1] Nef 1991, p. 100.
[2] Kripke 1982 [1972], p. 82-83 : « Il est faux que nous déterminons l'objet auquel nous faisons référence grâce à des propriétés qualitatives qui seraient à notre disposition et qui permettraient de singulariser l'objet en question. [...] En général, ce à quoi nous faisons référence dépend non seulement de ce que nous pensons nous-mêmes, mais des autres gens de la communauté, de l'histoire du chemin suivi par le nom pour nous atteindre, et ainsi de suite. C'est en suivant cette histoire qu'on parvient à la référence ».
[3] *Ibid.*, p. 95
[4] *Ibid.*, p. 124.
[5] Milner 1978, p. 333.
[6] *Ibid.*, p. 334, n. 1.

rend viable le pari de l'indexation, à quelques nuances près, que nous développerons dans le chapitre V consacré au descripteur : en effet, la notion de désignateur rigide ne recouvre qu'un type d'unité bien précis (essentiellement les noms propres et quelques cas de description définie) ; or l'indexation, si elle manipule bon nombre de noms propres, comme nous le montrerons ci-après, utilise surtout ce que les logiciens nomment les « désignateurs accidentels », soumis à la variabilité des univers de référence.

III.2.2 - PRATIQUES PROFESSIONNELLES ET USAGES DU NOM PROPRE

Au regard de ce que fait apparaître le principe de rigidité (un principe de désignation stable car indépendant des propriétés de l'objet que désigne un nom propre), il ne nous semble pas indifférent que nombre de pratiques recourent massivement à l'emploi de noms propres, la pratique d'indexation au premier chef.

A - Sur-représentativité du nom propre en indexation

Nous rendons compte ici d'observations tirées de l'expérimentation que nous avons menée[1].

Comparant les indexations réalisées par dix organismes documentaires sur un même numéro du journal *Le Monde*, nous avons relevé que la part des noms propres utilisés comme descripteurs oscille entre un tiers et deux tiers de l'ensemble des descripteurs utilisés[2]. On trouve principalement des noms de personnes, mais aussi des noms géographiques (pays, villes, régions) et des noms d'entreprises[3]. Outre ces données quantitatives qui indiquent déjà, comme le note Marandin et comme nous le développerons ultérieurement, une certaine « immédiateté » du choix lexical en faveur du nom propre[4], nous avons noté, dans les pratiques mêmes de l'indexation, une organisation privilégiant les accès par noms propres. Un cas exemplaire est illustré par le type d'indexation réalisé au centre de documentation de la *Fondation nationale des sciences politiques*. La classification utilisée donne la priorité à la localisation géographique, l'indication « thématique » étant seconde : « chaque fois que cela est possible, on classe d'abord sous le nom d'un pays, ou sous celui d'un ensemble géographique et seulement si aucune localisation n'est possible dans les rubriques générales[5] ».

Introduite sous forme de facettes et non plus d'indices de classification, l'entrée privilégiée par le nom propre est également à l'œuvre au centre de documentation du *Monde*. En effet, les descripteurs retenus doivent nécessairement figurer dans l'une des quatre rubriques suivantes : « territoires étrangers », « territoires français », « personnes physiques ou morales » et « auteurs cités ».

On notera également le caractère systématique de la mention de la localisation géographique dans les indexations réalisées à la *Documentation française* : le

---

[1] Le cadre de l'expérimentation est précisé dans l'annexe 1.
[2] Les résultats plus précis de cette étude sont présentés dans l'annexe 3.
[3] Pour des raisons de commodité, nous avons assimilé, dans notre typologie, les noms de manifestations culturelles (comme Le Festival d'Avignon, La Fureur de lire), les noms de partis politiques et les noms de syndicats dans une même catégorie, celle des « noms d'entreprises ».
[4] Marandin 1988, p. 79.
[5] Pour la journée d'indexation examinée, on a comptabilisé 5 cas sur 85 (environ 6 pour cent) de classifications échappant à ce primat de la localisation géographique.

champ « descripteurs géographiques », obligatoire, est distingué des champs d'indexation dits thématiques, où, par ailleurs, l'on trouve bon nombre de noms propres (de personnes, de partis politiques, etc.).

Cette sur-représentativité du nom propre en lieu et place de descripteurs « thématiques » nous paraît doublement significative : outre qu'elle confirme le projet de l'indexation de réaliser, par tous les moyens linguistiques possibles, une permanence référentielle, elle pointe également la faillite, ou plutôt, l'inadéquation des langages documentaires, qui tiennent qu'il est possible de traiter toutes les unités de langue comme des symboles sans signification. En effet, ce n'est pas parce que le descripteur doit, prioritairement, permettre de désigner, qu'il peut se passer de signifier.

Ce que souligne également la présence massive du nom propre en indexation, du moins dans le cadre de l'expérience réalisée, c'est le caractère discursif, énonciatif, de l'indexation : en effet, on peut rapprocher le phénomène observé d'un paramètre pragmatique dit « hétéro-facilitatif » à l'origine de comportements linguistiques destinés à éviter les ambiguïtés interprétatives[1] ; or, précisément, le nom propre, par opposition aux descriptions définies, neutralise ce type d'ambiguïtés.

Sur ce dernier point, notre expérience fait apparaître des cas d'indexation par nom propre qui ne semblent s'expliquer que par une volonté d'éviter toute ambiguïté interprétative.

Ainsi de l'indexation de deux interviews, l'une de Simone Veil, alors ministre des Affaires sociales, de la santé et de la ville[2], et l'autre de Jean Chrétien, premier ministre canadien[3]. Les deux interviews sont prioritairement, voire exclusivement[4], indexées par le nom propre de la personne interviewée, alors que Simone Veil est interrogée sur la question précise de la prévention du sida et non sur sa carrière politique ou personnelle. De même, Jean Chrétien s'exprime, dans cet article, non sur lui-même mais sur la présence canadienne dans l'ex-Yougoslavie.

Ces faits d'indexation nous semblent souligner deux points :

- d'une part que l'indexation ne consiste pas, dans la majeure partie des cas, à dire le contenu d'un texte : il s'agit avant tout d'essayer de le situer. En cela, les noms propres présentent des points d'ancrage forts, en termes spatio-temporels pour les noms géographiques, en termes de foyers énonciatifs pour les noms de personnes ;

---

[1] Apothéloz et Reichler-Béguelin 1995, p. 238 : « Il s'agit du contrôle par le locuteur du bon déroulement de l'acte référentiel, qui s'exprime en particulier dans l'anticipation des ambiguïtés référentielles ».

[2] *Le Monde* du 1/12/1994, p. 12.

[3] *Ibid.*, p. 6.

[4] Par exemple, concernant l'entretien de Simone Veil : sur les six organismes documentaires qui ont sélectionné cette source, deux ne l'indexent que par le nom propre « Veil Simone » ; deux l'indexent par le nom propre et y ajoutent des noms communs ; un des organismes documentaires qui ne dispose pas, dans la liste des descripteurs autorisés, de noms propres de personnes (accessibles à la recherche par d'autres biais) effectue une indexation qui pointe sur la personnalité de l'interviewée : « structure du gouvernement / personnage » ; enfin, le dernier organisme documentaire à avoir sélectionné cet article hésitait entre une indexation « biographique » [par nom propre] jugée non satisfaisante et une indexation « thématique » qui semblait trop floue (indexation par le terme « sida »).

- d'autre part, et la remarque d'un indexeur est sur ce point éclairante[1], le recours au nom propre apparaît comme le moyen de ne pas prendre de « risque interprétatif » ; l'indexation par nom propre permet à coup sûr de retrouver un document mais elle dispose ce faisant d'accès très sélectif : il faut *connaître* préalablement le nom propre, on ne peut pas le « trouver ».

Les noms propres en indexation, par leur nombre d'une part et par leurs utilisations – souvent inattendues – d'autre part, ne peuvent manquer d'attirer l'attention de l'analyste ; cependant, nous n'avons pas trouvé, dans la littérature consultée, de remarques ni d'explications de ce point.

Dans le cadre de notre analyse, la présence massive des noms propres en indexation, qui mériterait une réflexion plus approfondie, se comprend comme un moyen, pour une pratique, de stabiliser l'« instabilité constitutive des objets de discours[2] » afin de présenter une version stable du monde.

D'autres pratiques recourent à un semblable procédé de stabilisation.

B - Attractivité du nom propre dans les pratiques

L'immédiateté du choix lexical en faveur du nom propre se marque, de façon générale, dans les stratégies de désignation adoptées par les sujets parlants, qu'ils soient ou non en situation professionnelle.

Sur ce point, Marandin relève, dans son analyse du thème de discours, que la détermination du nom d'un thème n'est pas insensible à l'attractivité du nom propre[3]. Il propose une explication en termes de « cause ultime », qu'il reprend de Ricœur [1977] : le nom propre « révèle le caractère fini de l'explication », le nom propre sature une interprétation, pourrait-on dire.

En matière de pratique professionnelle, Beaulieu [1995] remarque : « Les mass-media et, en particulier, les concepteurs de périodiques culturels ont abondamment recours au nom propre pour stimuler la reconnaissance d'une information *a priori*, première phase du processus de la connaissance. Mais que se passe-t-il ensuite ? De quelle manière participe-t-il à l'élaboration des idées en s'insérant dans la consécution des inférences ? En approfondissant une information déjà contenu dans les prémisses d'un raisonnement ou en produisant quelque chose qui ne s'y trouve aucunement impliqué ? Il y a là un choix qui, d'après nous, explique une partie des écarts entre le savoir des individus.[4] »

Nous reprenons l'intégralité des interrogations de Beaulieu sur le sur-emploi des noms propres dans certaines pratiques dans la mesure où elles soulignent une limite spécifique à l'utilisation du nom propre : au-delà de la reconnaissance d'un référent, que permet le nom propre d'un point de vue interprétatif ? S'il est saturé de ce point de vue, peut-il permettre de construire plusieurs objets de discours ?

---

[1] Un commentaire était joint à l'indexation de l'article de Simone Veil : « Avant, on aurait indexé en biographie, mais maintenant on essaie de privilégier le contenu » : l'indexation par nom propre apparaît ici comme ce qui permet de ne pas s'exprimer sur le « contenu ».
[2] Dubois et Mondada 1995, p. 273.
[3] Marandin 1988 p. 79 (et note 17) : « /Lancelot/ constitue un choix plus immédiat que /la vengeance des licornes/ ».
[4] Beaulieu 1995, p. 41.

Nous examinerons de près ce point dans le chapitre V : si le nom propre semble se présenter comme le candidat descripteur idéal, en ce qu'il permet d'assurer une stabilité de la référence, il apparaît bien vite comme manquant cruellement de « plasticité » discursive ; or c'est plus cette plasticité-là que l'idéal d'une stabilité référentielle qui se révèle importante en indexation.

Nous avons proposé de voir, dans l'utilisation des noms propres en indexation, un moyen qu'utilise la pratique pour assurer une stabilité référentielle : nous avons dégager ce processus de stabilisation en nous appuyant sur le modèle de la rigidité établi par Kripke.

Ce modèle montre les enjeux, pour la construction de la référence en indexation, de l'emploi de telle ou telle unité. Le recours aux noms propres et/ou aux noms communs n'est pas indifférent parce que leur pouvoir référentiel n'est pas le même : si le nom commun réfère en discours à des objets différents, c'est sur la base de sa signification lexicale ; le nom propre réfère, lui, à un individu particulier de façon régulière sur la base d'un acte d'énonciation. À confondre les deux types d'unité, on en vient à confondre aussi les deux types de fonctionnement et à postuler ainsi que l'on peut « rebaptiser » les mots de la langue naturelle par des descripteurs de tout type linguistique.

IV - Conclusion du chapitre

La question de la référence nous est apparue centrale en indexation, notamment sous deux aspects :

- celui de la stabilité référentielle d'une part, puisque sa recherche constitue l'objectif premier et explicite de l'indexation ;

- celui des objets référentiels visés d'autre part, dans la mesure où il concerne la cible matérielle de l'indexation.

Sur ces deux points, le modèle réaliste sous-jacent à l'approche classique de la référence ne nous fournit guère d'éléments : la stabilité référentielle y apparaît sous la forme d'un axiome, en cela peu apte à rendre compte des faits de variation observables dans les pratiques. Les objets de l'indexation, le document notamment, semblent être exclus du champ des interrogations, en raison même du présupposé de leur préexistence.

Ces deux aspects de la référence en indexation restent difficiles à appréhender dans le cadre du modèle réaliste. On peut montrer en effet qu'ils engagent une réflexion plus globale, d'une part sur le rapport entre le sens et la référence, et d'autre part sur la relation entre objets mondains et objets textuels, qui ne se comprend que dans un modèle non réaliste de la référence tel qu'il se donne dans le champ linguistique.

Il nous est apparu, à ce titre, nécessaire de rappeler les grandes lignes de la problématique de la référence en linguistique : les distinctions entre types de référent comme entre acte de référence et propriété référentielle, mais aussi entre signification lexicale et construction référentielle.

Sous le nouvel éclairage retenu, nous avons repris les deux principaux problèmes de référence que nous avions identifiés :

- l'examen concernant la construction du référent-document en indexation nous a amené à concevoir l'opération d'indexation comme un double processus de contextualisation, comprenant la phase dite de sélection documentaire, assimilée alors à une étape de décontextualisation, à laquelle répond une étape de recontextualisation, au terme de laquelle une source devient un document ;

- l'étude de l'effet de stabilité référentielle recherché en indexation nous a conduit à disposer, au cœur de la construction de la référence documentaire, le principe de rigidité tel que Kripke a pu le concevoir. Sous cet angle, il apparaît que le descripteur révèle un comportement proche de celui du désignateur rigide, la première manifestation notable en étant l'utilisation massive du nom propre en indexation.

Chemin faisant, est apparue une dimension – celle du discours – qui, pour être nécessaire à la compréhension de l'acte de référenciation en indexation, n'en est pas moins absente des approches classiques. C'est ainsi que si le chapitre II, consacré au lexique, nous amenait à étudier la question de la référence, ce chapitre III nous conduit à examiner la dimension du discours en indexation : nous effectuons là un saut, car notre approche de l'indexation s'éloigne alors radicalement du discours classique. Il nous semble nécessaire de franchir le pas. En effet, la constitution des fondements théoriques de l'indexation, telle que nous l'entendons dans cette recherche, risque sinon de rester incomplète. Certes, nous avons proposé, dans le paragraphe III de ce chapitre, un modèle d'utilisation de l'approche non réaliste de la référence en indexation, mais il nous manque encore beaucoup d'éléments pour que les remarques proposées puissent constituer des fondements théoriques : en quoi un document constitue-t-il un « énoncé » ? En quoi l'indexation réalise-t-elle une « énonciation » ? Comment se définissent, dans ce cadre, le rôle et la morphologie du descripteur ?

En faisant le bilan des problèmes théoriques de l'indexation que nous avons pu dégager par notre étude du lexique et de la référence en indexation, nous pourrons proposer un modèle d'utilisation de la langue qui repose sur un modèle de fonctionnement de la langue valide du point de vue de la théorie linguistique : en cela, les réponses que nous proposerons, notamment aux trois questions précédemment posées, pourront prétendre à une certaine généralité.

C'est à la mise au point d'un modèle d'utilisation de la langue en indexation que nous consacrons la seconde partie de cette étude.

CONCLUSION DE LA PREMIÈRE PARTIE

Notre étude de deux problèmes théoriques de l'indexation a permis de montrer que la pratique d'indexation, pour peu que l'on reformule les éléments par lesquels elle est classiquement décrite, se fonde, au moins en partie, sur des propriétés de la langue et du langage :

- l'étude du lexique en indexation (du rôle des mots) a montré que les concepts, proposés par la linguistique contemporaine, de « signification lexicale » et de « synonymie référentielle », étaient à même d'expliquer sur quoi reposent une partie des faits d'indexation. Ces deux propriétés des unités lexicales permettent en effet de comprendre : pourquoi l'indexation peut utiliser des formes linguistiques isolées et pourquoi ces formes linguistiques sont nécessairement de type nominal (nom commun ou nom propre) ; pourquoi une même forme linguistique peut renvoyer à différents types d'objets. De ce point de vue, on pourrait considérer que les concepts linguistiques de signification lexicale et de synonymie référentielle constituent des fondements théoriques de l'indexation. Mais on remarque aussitôt que ces deux propriétés linguistiques sont utilisées, en indexation, sans que soient distingués les différents niveaux où elles sont à l'œuvre : langue et discours. Dès lors, ces propriétés linguistiques, si elles peuvent constituer des fondements théoriques de l'indexation, doivent être resituées dans le cadre d'un modèle d'utilisation de la langue qui permette de les rendre réellement « actives ». Nous avons en ce sens proposé une première piste dans le chapitre II ;

- l'étude de la référence a permis de dégager une propriété du langage lui-même à l'œuvre en indexation. De même que le langage parle du « réel », l'indexation parle aussi du « réel » ; elle manipule cependant des objets du monde réel au statut sémiotique particulier : les textes. On comprend du coup pourquoi l'indexation ne peut se passer des mots : quels sont les autres « instruments » qu'elle pourrait utiliser pour parler du monde ? On mesure aussi, dès lors que l'on considère l'indexation sous l'angle de son rapport avec le réel, la complexité des relations qu'elle institue : quel est le réel de l'indexation ? Le monde des objets, le « quasi-monde des textes » ? L'examen des objets manipulés montre que l'indexation, pour pouvoir parler des objets du monde, construit ses propres objets à partir du monde des textes. Dès lors la problématique de l'indexation s'élargit à la construction du document : l'indexation n'est plus seulement une opération qui attribue des mots à des textes mais aussi, et surtout, une opération qui constitue un monde de

documents. À ce titre, on a proposé une schématisation de l'indexation qui puisse rendre compte de la construction des documents au sein du processus de l'indexation elle-même (*supra* figure 3).

Cet ensemble de propriétés linguistiques disparates (signification lexicale, synonymie référentielle, construction de la référence), à l'œuvre dans l'indexation, s'appréhende de façon unifiée par le biais de la notion de thème discursif dégagée d'un point de vue linguistique : il apparaît en effet que le principe de l'indexation repose sur le fonctionnement linguistique du thème de discours.

Le point de vue linguistique sur le thème montre que la construction du thème discursif est un fait d'interprétation : un thème n'existe pas en soi dans un texte ou plusieurs textes, on ne peut donc l'extraire. Un thème ne se constitue qu'au cours d'une lecture qui met en jeu plusieurs textes. Dès lors, si l'indexation consiste à permettre de dégager des thèmes, sa problématique se laisse redéfinir :

- du point de vue des documents : comment l'indexation peut-elle construire un monde des documents qui autorise une lecture interdiscursive ?

- du point de vue des descripteurs : quel type de forme linguistique l'indexation doit-elle proposer comme accès pour permettre un tel parcours intertextuel ?

Ces deux questions se situent au cœur du modèle de l'utilisation de la langue que nous proposons dans la seconde partie de cette étude.

Deuxième partie

Contribution aux fondements théoriques de l'indexation

Nous avons précédemment proposé de schématiser le mécanisme de l'indexation de la façon suivante :

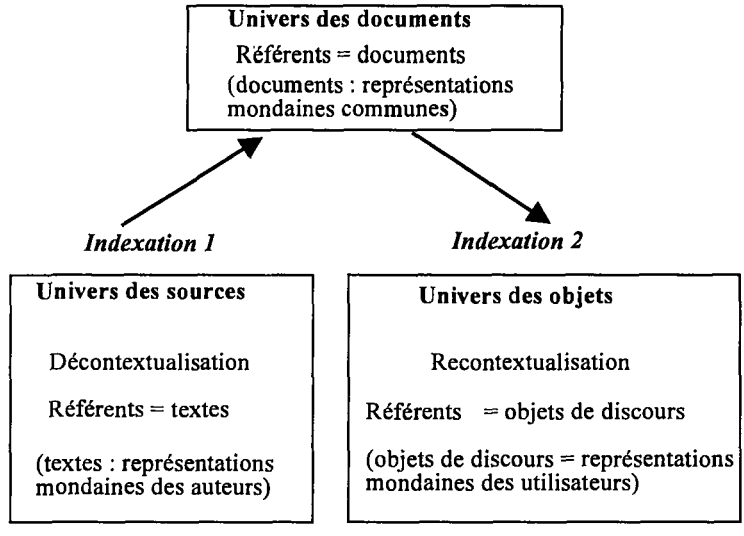


Figure 3 – Mécanisme « à double détente » de l'indexation

Ce mécanisme distingue deux moments dans l'indexation qui ne mettent pas en jeu les mêmes mécanismes référentiels ni exactement le même type d'acteurs :

- l'indexation 1, qui réalise une décontextualisation des sources, correspond à l'acte de référence (l'acte de discrétisation du quasi-monde des textes) réalisé par les indexeurs : c'est le moment de la construction du monde de référence, du monde des documents. La problématique de l'indexation est alors celle du

type de sélection et du type de regroupement des sources qu'il faut réaliser pour rendre une collection documentaire exploitable ;

- l'indexation 2 correspond à une recontextualisation des sources : c'est le moment de la lecture des documents dans le cadre d'une problématique spécifique, propre à un utilisateur (c'est là qu'apparaît l'« effet stabilisateur des pratiques »). Cette phase met en jeu les utilisateurs mais aussi les indexeurs : ils participent à la recontextualisation en disposant des « accès » aux sources susceptibles de permettre la construction de thèmes de discours. C'est ici les propriétés référentielles des descripteurs qui concernent la problématique de l'indexation.

Alors que l'indexation est habituellement envisagée sous l'angle de la seule attribution de descripteurs (indexation 2 dans notre schéma), nous proposons de la considérer de façon plus large comme un processus qui organise un univers référentiel spécifique. L'indexation ne se laisse donc plus voir sous la seule dimension lexicale ; elle met aussi en jeu, comme l'a montré Michel Le Guern [1984 et 1991a], une dimension discursive.

L'introduction de cette dimension du discours se fonde sur l'approche linguistique du thème discursif, dont nous avons posé qu'elle pouvait permettre de fonder l'indexation d'un point de vue théorique. Nous proposerons donc un modèle d'utilisation de la langue en indexation qui rende possible la construction de thèmes de discours telle qu'une approche linguistique la décrit. En ce sens, cette seconde partie s'organisera de la façon suivante :

- dans le chapitre IV, nous définirons ce que peut être l'« interdiscours » en indexation ; nous traiterons alors l'indexation 1 de la figure 3 ci-dessus : l'indexation comme processus ;

- dans le chapitre V, nous préciserons le type d'unités linguistiques susceptibles de permettre la construction du thème discursif ; c'est alors l'indexation 2 qui nous occupera essentiellement : l'indexation comme résultat (le descripteur).

CHAPITRE IV
LA DIMENSION DISCURSIVE DE L'INDEXATION

L'objectif de ce chapitre est de définir le *processus* de l'indexation dans le cadre d'un modèle de fonctionnement de la langue, ou, plus précisément, dans le cadre d'une approche linguistique du lexique et de la référence.

Trouvant ses fondements dans un modèle linguistique du lexique et de la référence, notre approche de l'indexation en termes de discours ne sera pas elle-même de nature strictement linguistique. Notamment, la notion de discours, telle qu'elle sera traitée dans ce chapitre, se laissera approcher dans le cadre plus large d'une analyse sémiotique : sauf indication contraire, on entendra ici par discours un « ensemble d'énoncés et/ou de textes, *possédant une organisation* thématique, normative, structurale[1] ». C'est principalement sous l'angle de l'organisation de textes que nous étudierons la dimension discursive de l'indexation : à ce titre, le discours documentaire se conçoit comme un espace d'organisation spécifique des documents obtenu par le biais d'« une transformation discursive réglée ».

Nous nous attacherons dans ce chapitre à dégager les *principes* qui sous-tendent les deux actes de référenciation réalisés en indexation[2] :

- quels sont les principes qui dictent l'acte de discrétisation sur les sources ?
- quels sont les principes qui guident l'acte de stabilisation de l'espace documentaire ?

Ces deux types de principe seront analysés en termes de *stratégie* : on distinguera, d'une part, l'indexation en tant que stratégie d'exploration des sources et, d'autre part, l'indexation en tant que stratégie d'exposition des documents.

La dimension discursive de l'indexation est particulièrement absente des discours classiques : comme nous l'avons vu dans le chapitre II, c'est la dimension lexicale qui prédomine. De même, on s'interroge généralement peu sur les principes de

---

[1] Souchard 1989, p. 258 (c'est nous qui soulignons).
[2] Voir chapitre III.

sélection des sources en indexation, même si leur importance a pu être relevée[1]. Enfin, la notion de stratégie d'exposition des documents reste en général appréhendée par le seul biais des langages documentaires.

Il nous faut donc substituer, dans un premier temps, à la notion de langage documentaire, la notion de discours documentaire, de façon à aborder, dans un second temps, d'une part, la constitution de ce discours documentaire (stratégie d'exploration des sources) et, d'autre part, les conditions de son utilisation (stratégie d'exposition des documents).

Nous procéderons donc dans ce chapitre en trois étapes :

- nous tâcherons d'abord de définir ce que l'on peut entendre par discours documentaire, notion que l'on substituera à celle de langage documentaire. En distinguant langage et discours documentaires, nous cherchons à distinguer les mots (langage documentaire) des utilisations qui en sont faites (dans un espace de discours). Pour ce faire, nous examinerons comment l'approche linguistique de pratiques comme la terminologie et la vulgarisation scientifique a appréhendé cette question : sur ce point, les problématiques de l'indexation rejoignent celles d'autres pratiques ;

- nous nous intéresserons ensuite à l'espace de discours que construit l'indexation à partir des sources, en traitant la question suivante : sur quels principes l'indexation sélectionne-t-elle ses sources ? Nous nous appuierons sur les notions de « système-archive » et de « formation discursive » proposées par Foucault ;

- nous nous pencherons, pour finir, sur l'articulation que l'indexation met en place entre les trois espaces dans lesquels elle évolue : espace des auteurs (sources), espace des indexeurs (documents), espace des utilisateurs (objets de discours). Cette articulation sera pensée par le biais de la notion de « monde possible » proposée par Kripke.

I - Langage ou discours documentaire ?

Nous situerons, dans un premier temps, la question que nous posons (« langage ou discours documentaire ? ») dans le cadre plus large de la réflexion des pratiques sur elles-mêmes.

Nous aborderons ensuite l'approche linguistique dont ont fait récemment l'objet la terminologie et la vulgarisation scientifique, deux pratiques professionnelles qui nous paraissent pouvoir être fructueusement rapprochées des pratiques documentaires.

---

[1] Par Bertrand-Gastaldi 1986, par exemple, p. 4 : « Nous n'étudierons pas le sous-système de collecte qui, pourtant, constitue le premier filtre important grâce à la sélection d'un nombre réduit de domaines du savoir, de publications, de supports, en fonction des objectifs poursuivis ».

I.1 - Notions de langage et de discours dans les pratiques : enjeux

Avancer la notion de discours documentaire peut sembler incongru dans la mesure où l'indexation ne crée pas de textes à proprement parler. Ce que nous cherchons à montrer en opposant langage documentaire à discours documentaire, c'est que l'indexation n'utilise pas un langage particulier, ce que laisse entendre la notion de langage documentaire, mais qu'elle fait une utilisation documentaire du langage (ou plutôt des productions du langage), ce que voudrait faire entendre la notion de discours documentaire.

Il nous semble tout à fait important de noter que, sur ce point, la pratique d'indexation n'est pas la seule pratique professionnelle à difficilement intégrer la dimension discursive dont elle relève. Nous avons retenu, à la fois pour exemples et pour guides, les pratiques de la terminologie et de la vulgarisation scientifique qui ont récemment connu un renouvellement de leurs problématiques dans un sens proche de celui que nous voudrions introduire dans cette recherche. Aussi bien la terminologie que la vulgarisation scientifique constituent des pratiques qui articulent des ensembles de mots (en l'occurrence des terminologies) et des ensembles de textes (textes spécialisés ou textes vulgarisés). La perception des pratiques par elles-mêmes tend à focaliser l'attention sur les mots eux-mêmes ; la perception linguistique des pratiques tend, elle, à distinguer les mots d'une part et leur rôle dans un type de discours d'autre part. C'est cette distinction que nous cherchons à mettre au jour dans la pratique d'indexation.

Sur ce point, il ne nous semble pas que les critiques courantes des langages documentaires permettent réellement de mettre à distance la dimension lexicale en indexation. On ne compte plus les condamnations des thésaurus et autres langages « autoritaires[1] » et nombre de systèmes d'indexation automatique[2] fondent leur argumentaire de vente sur le « texte intégral », autrement dit sur l'absence de langage contrôlé. Cependant, ces critiques ne remettent pas en cause la dimension lexicale de l'indexation ; elles discutent simplement la forme, plus ou moins artificielle et plus ou moins figée, des descripteurs.

Ce que nous voudrions montrer c'est que, si « lexique documentaire » il y a en indexation (sous forme de liste de descripteurs), il est à comprendre dans un cadre discursif, dans un cadre où les documents peuvent se constituer comme des discours qui s'interpellent *via* des mots. C'est dans ce cadre discursif que les descripteurs peuvent alors être approchés (Chapitre V).

---

[1] Voir, par exemple, Turner 1990.

[2] Essentiellement ceux qui exploitent une liste inverse de chaînes de caractères. L'indexation est entendue dans ce cas dans une acception informatique, qui repose sur la notion informatique d'« index » : « L'index d'une banque de données texte intégral est le plus souvent représenté sous la forme d'un fichier inverse. Dans un fichier inverse, un enregistrement est créé pour chaque terme apparaissant dans la collection des documents à indexer. À chacun de ces termes est associée une liste de couples de la forme (identificateur de fichier, position dans le fichier). Chaque élément de cette liste correspond à une occurrence du terme dans la collection des documents. Si le mot *avion* apparaît dans le fichier 12 à la position 16 et dans le fichier 15 à la position 18, le fichier inverse contient une entrée de la forme : *avion* (12 16) (15 18) ». Rôle 1993, p. 137.

I.2 - Évolution des problématiques en terminologie : émergence de la notion d'usage professionnel de la langue

Il nous semble que la question « langage ou discours documentaire » s'éclaire comparée à la problématique « langue ou discours de spécialité » qui caractérise les débats en terminologie (I.2.1). Par ce nouvel éclairage, nous pourrons restituer au discours documentaire ses deux dimensions discursives – textuelle et énonciative – particulièrement peu visibles dans les discours normatifs (I.2.2).

I.2.1 - PROBLÉMATIQUES DE LA TERMINOLOGIE

La terminologie est avant tout une pratique professionnelle (historiquement un « art ») qui cherche à se constituer, depuis peu, en discipline scientifique spécifique, à la recherche de fondements théoriques, entre autres linguistiques[1].

L'émergence d'une approche théorique s'est faite en terminologie sur la base de reformulations importantes. L'enjeu a notamment consisté à souligner que la terminologie ne pouvait se concevoir sous l'angle du seul lexique, comme « un lexique spécialisé » ou encore comme une « langue spécialisée », mais qu'elle devait se comprendre sous l'angle du discours[2]. Dès lors, c'est la notion même de terme* qui s'est trouvée redéfinie. Elle ne correspond plus à un type de mot particulier, caractérisé par sa monoréférentialité (en opposition aux mots non spécialisés qui peuvent, eux, être polyréférentiels) ; elle correspond plutôt à un usage particulier d'un certain type de mots. La notion de terme se laisse alors capter à plusieurs niveaux, à la fois à un niveau linguistique (en termes de propriétés) et à un niveau extra-linguistique (en termes d'usages sociaux, professionnels).

En tant que pratique définie par des normes, la terminologie[3] distingue les notions de « langues de spécialité » et de « terminologie[4] », celle-ci étant intégrée à celle-là : « On entend par langue de spécialité "un sous-système linguistique qui utilise une terminologie et d'autres moyens linguistiques et qui vise la non-ambiguïté de la communication dans un domaine particulier".[5] »

D'emblée, notamment d'un point de vue linguistique, la notion de « langue de spécialité » apparaît problématique, dans la mesure où elle ne saurait être assimilée à un type de langue, à un dialecte : la langue est, en terminologie, présente telle qu'en elle-même. Elle est, en revanche, comme dans l'indexation, utilisée dans une perspective particulière, professionnelle : « Le français de l'automobile est l'*usage du français* pour rendre compte de connaissances en matière d'automobile. [...] La langue spécialisée est d'abord une *langue en situation* d'emploi professionnel (une "langue en spécialité" comme dit l'école de Prague). C'est la langue elle-même

---

[1] Un aperçu historique se trouve dans Rey 1992 ; de façon plus précise, les prémices de la réflexion terminologique sont identifiées dans Mustafa-Elhadi 1989. Lerat 1995 propose une synthèse des approches théoriques.
[2] Voir Le Guern 1989 notamment.
[3] L'ISO définit la terminologie comme l'« étude scientifique des notions et des termes en usage dans les langues de spécialités », ISO 1087 (1990).
[4] « Terminologie » est à entendre ici « comme un ensemble d'expressions dénommant dans une langue naturelle des notions relevant d'un domaine de connaissances fortement thématisé », Lerat 1995, p. 20.
[5] ISO 1087 1990 cité *in* Lerat 1995, p. 17.

(comme système autonome) mais au service d'une fonction majeure : la transmission des connaissances.[1] »

Ce changement de perspective, de « langue spécialisée » à « usage spécialisé », conduit les théoriciens à privilégier la notion de « discours de spécialité », rejetant comme impropre la dénomination de « langue de spécialité[2] ».

Avec la notion de discours de spécialité se repositionne celle de « terminologie » (en tant que liste de termes). Elle n'est plus dans un rapport hiérarchique d'inclusion (comme l'indique la norme ISO), mais dans un rapport dynamique de définition avec les textes spécialisés : « Les textes savants saisissent et expriment le contenu savant, dont les unités sémantiques dominantes sont les termes. Le terme est une unité lexicale – ou son acception – définie dans les textes savants, où apparaissent ses occurrences intégrées dans le tissu du texte. [...] Les termes ne sont pas seulement des éléments du système, mais des occurrences dans les textes savants.[3] »

Mais les textes spécialisés ne se contentent pas de définir les termes ; ils les constituent également en tant que tels, dans la mesure où « ce n'est que l'emploi répété du terme dans les textes différents qui mène, au fur et à mesure, à son implantation, à sa vraie terminologisation, à son passage du texte au système. Ce sont les textes spécialisés qui décident du statut systémique d'un néologisme.[4] »

Ainsi l'existence des termes apparaît-elle complètement déterminée par des textes dont cependant le caractère « spécialisé », « savant », est lui-même fortement déterminé par la présence des termes ...

La dimension textuelle ne permet donc pas de constituer à elle seule un « discours spécialisé » ; intervient également une dimension énonciative, celle des locuteurs : « En fait, la dichotomie langue générale/langues de spécialité se fait à partir de l'opposition expérience partagée par l'ensemble des locuteurs/expérience partagée par des sous-ensembles de locuteurs.[5] »

Ainsi les notions de discours spécialisé et de terminologie semblent-elles se constituer autour de l'idée d'espace d'expérience partagée.

Rien de moins hypothétique, bien entendu, qu'un découpage en cercles, plus ou moins larges, de locuteurs ; c'est pourtant bien l'idée d'une convention entre locuteurs qui légitime et fonde une utilisation partielle et spécialisée de la langue en terminologie.

C'est ainsi que la notion d'espace discursif a été, en terminologie, peu à peu explicitement posée comme centrale par les chercheurs du domaine[6]. C'est elle qui

---

[1] Lerat 1995, p. 21 (c'est nous qui soulignons).
[2] Voir Rey par exemple : « Il n'y a pas à proprement parler de "langue" mais des "vocabulaires", des "usagers" et des "discours de spécialité" », ou encore Quémada : « La linguistique descriptive [...] condamne les désignations de "langue" technique et scientifique qui sont également impropres ». Citations extraites de Kocourek 1991a, p. 15.
[3] Kocourek 1991b, respectivement p. 71 et p. 72.
[4] *Ibid.*, p. 73.
[5] Portelance 1989, p. 402.
[6] Voir par exemple, Boutayeb 1993, p. 14 : « Dans les unités terminologiques, l'expression (c'est-à-dire la forme) résulte d'une convention qui est le résultat de l'accord des spécialistes

permet de concevoir la notion de concept* à laquelle se réfère la pratique terminologique : il ne s'agit plus de concepts liés aux mots eux-mêmes mais de concepts liés aux objets scientifiques manipulés par un nombre restreint de locuteurs. Comme l'indique Mortureux, « ce type de discours [le discours spécialisé] s'échange en général à l'intérieur d'un cercle de spécialistes, et par conséquent les questions que peut soulever la réception des termes sont intimement liées aux concepts et relèvent plus de la démarche scientifique que de la compréhension strictement linguistique[1] ».

Une approche linguistique de la pratique terminologique permet de mettre en valeur les deux facettes de la dimension discursive des terminologies (dimension textuelle et dimension énonciative), rendant délicate toute approche qui considérerait les termes comme des types de mot particuliers :

- *d'une part, elle montre que les termes des terminologies sont indissociablement liés aux discours qui les instituent ;*
- *d'autre part, elle souligne que les termes ne sont pertinents[2] que dans le cadre restreint d'une pratique technique ou scientifique donnée.*

Si la terminologie vise « la non-ambiguïté de la communication », comme l'indique l'ISO, ce ne peut être que dans un cadre doublement restreint et par un ensemble de textes (notion de discours spécialisé) et par un ensemble de locuteurs (notion de cercles étroits de spécialistes). Créant un espace de textes et dessinant un espace de locuteurs, la terminologie se conçoit alors comme une pratique qui spécifie un usage « spécialisé » de la langue et non plus comme une pratique qui construit des « langages de spécialité » ou des « langues spécialisées ». C'est ainsi que la pratique terminologique consiste essentiellement[3] à construire des corpus de textes « spécialisés » et à établir, à partir d'eux, le mode d'emploi de certains mots, les termes, et ce, à l'intention non du « grand public » mais « des locuteurs spécialistes du domaine d'origine ou d'un domaine connexe[4] ».

En quoi les problématiques de la terminologie et le regard croisé, pratique et théorique, dont elles ont fait l'objet peuvent-elles nous permettre d'approcher la dimension discursive de l'indexation ?

I.2.2 - RAPPROCHEMENT DE DEUX DISCIPLINES

On notera tout d'abord que le rapprochement entre pratique documentaire et pratique terminologique n'est pas nouveau : il est même en plein essor.

Michel Le Guern et les membres du groupe SYDO[5] ont, dès les années 80, suscité l'attention des professionnels : les propriétés du terme comme unité de discours méritent d'être rapprochées des fonctionnalités attendues du descripteur.

---

d'un domaine scientifique donné. Le terme s'impose par une pratique unifiée dans un milieu d'experts ».

[1] Mortureux 1995, p. 23.

[2] Les termes ont pour fonction de réaliser une désignation référentielle univoque (un terme = un concept ; un concept = un terme). On peut dire que leur pertinence est liée à la réalisation de cette fonction.

[3] Mais pas uniquement, voir sur ce point, dans le glossaire, l'entrée « analyse terminologique ».

[4] Mortureux 1995, p. 23.

[5] Pour les textes fondateurs, voir Le Guern 1984, Bouché 1989, Metzger 1988.

Ce n'est que dix ans après, et dans des termes sensiblement différents, que la pratique terminologique est rapprochée de la pratique d'indexation par les professionnels eux-mêmes. Les années 1990 ont vu en effet se multiplier les rencontres « terminologie/documentation[1] ». Parallèlement, à la même époque, les concepteurs de systèmes d'indexation automatique ou automatisée[2] renouvellent leur argument de vente : l'utilisation des terminologies pour l'indexation et la recherche documentaires permet d'assurer des taux de rappel* et de précision* bien supérieurs à ceux obtenus par l'utilisation des seuls thésaurus[3].

Ce qui retient l'attention, semble-t-il, des professionnels de l'information, c'est le projet de la terminologie d'établir une « communication non ambiguë » *via* les termes d'une terminologie. Les professionnels fondent leur espoir sur cette caractéristique du terme sans toujours observer que celle-ci suppose à la fois un domaine restreint de textes et un cercle circonscrit de locuteurs. Dans une telle perspective d'utilisation des terminologies en indexation, il s'agit toujours de recourir à un lexique, ici constitué de termes, pour réaliser l'indexation : on ne sort pas à proprement parler du domaine lexical et on ne prend pas en compte la dimension discursive des termes[4].

Si la pratique de l'indexation peut être, utilement, nous semble-t-il, rapprochée de la pratique terminologique, elle ne peut cependant complètement s'y fondre. Les raisons sont nombreuses qui peuvent permettre d'argumenter ce point. Nous n'en retiendrons qu'une dans cette recherche : alors que la pratique terminologique vise à établir une communication entre spécialistes, une communication « spécialisée », la pratique d'indexation vise à établir une communication plus large entre non-spécialistes (ou du moins pas nécessairement spécialistes) au sein de cercles de locuteurs appartenant à des horizons divers : on parlera alors de communication « vulgarisante ».

Si, à un niveau qu'il conviendra de préciser, les deux pratiques terminologique et documentaire peuvent être rapprochées, il convient d'abord de montrer qu'elles mettent en jeu des problématiques similaires.

L'approche linguistique de la terminologie a permis de passer d'une perception de la terminologie comme « langue de spécialité » à une appréhension en termes de « langue en spécialité », une langue en situation d'usage professionnel. Un tel déplacement nous semble pouvoir être réalisé en indexation. On cherchera à

---

[1] Par exemple, en 1993, le département *Formation continue* de l'Institut d'Études Politiques de Paris organise un séminaire de trois jours, intitulé « Terminologie et documentation » ; l'expérience est renouvelée depuis tous les ans. La Direction de l'information scientifique et technique du ministère de l'Enseignement supérieur et de la recherche met l'accent sur ce thème en 1994 en organisant la première journée « Terminologie et Information ». Dernier exemple en date : l'ADBS (Association des professionnels de l'information et de la documentation) a organisé, dans sa journée d'étude du 19 décembre 1996 consacrée aux « Outils linguistiques et [aux] nouvelles technologies », une table ronde sur les synergies entre terminologie et thésaurus.

[2] Dans le sens donné en note 2, p. 169.

[3] Voir par exemple le colloque RIAO (Recherche d'information assistée par ordinateur) de 1994 qui relève dans son compte rendu, p. 13 : « Les descriptions formelles du terme que nous fournissent les terminologues peuvent être utilisées pour automatiser, avec beaucoup de succès, la reconnaissance des descripteurs » ; des systèmes automatisés sont présentés dans le même volume.

[4] Nous proposons dans le chapitre V une tout autre approche des termes en indexation qui reprend les présupposés établis par Le Guern et les membres de l'équipe SYDO.

montrer que la notion de langage documentaire est inadéquate et qu'il faut lui substituer la notion de discours documentaire, au sens d'usage documentaire des mots. Pour cela, nous nous appuyons sur les deux aspects de la dimension discursive mise en valeur par l'approche linguistique de la terminologie : comment se manifestent en indexation la dimension textuelle et la dimension énonciative des descripteurs ?

A - Dimension textuelle des descripteurs

De la même façon que les termes sont créés par et dans les discours spécialisés, on peut dire que les descripteurs sont créés par et dans des discours que l'on qualifiera de documentaires[1].

Le discours normatif de l'indexation, qui met en avant la notion de langage documentaire, ne dit rien de l'origine textuelle des mots retenus pour figurer dans ce langage au titre de descripteurs[2]. Or un langage documentaire est toujours constitué de mots extraits d'un corpus de textes, de « mots de discours ». Nous illustrerons cet aspect des langages documentaires par quelques exemples concernant la constitution (A1) et la mise à jour (A2) de thésaurus.

A1 - Les normes[3] et les manuels[4] indiquent deux méthodes pour collecter les termes[5] d'un thésaurus, qui toutes d'eux consistent en une extraction de mots du discours.

Dans la première méthode, dite inductive (ou synthétique), les termes sont extraits de référentiels existants (dictionnaires spécialisés, nomenclatures, terminologies, tables de matière, classifications, etc.) ; dans la seconde méthode, dite déductive (ou encore analytique), les termes sont extraits d'un corpus de documents (de 1 000 à 5 000 documents). Généralement, il est conseillé d'employer les deux méthodes et de recourir à des « spécialistes » pour déterminer les termes importants d'un « domaine ». Sur ce plan, on le voit, la collecte des descripteurs se différencie peu d'une collecte de termes que réaliserait un terminologue. Cependant, si un terminologue a pour charge de référencer avec précision la source dont il extrait un terme, le compilateur d'un thésaurus, lui, s'affranchit plus radicalement de l'origine textuelle des termes qu'il retient.

A2 - C'est sur le même mode d'extraction de mots de discours que s'effectue la mise à jour des thésaurus. Un terme est repéré dans la littérature comme susceptible de désigner un « concept » nouveau ; il est alors constitué comme candidat-descripteur généralement jusqu'à ce que le nombre de ses apparitions dans les documents justifie son insertion dans le thésaurus. Un terme repéré dans les

---

[1] Dans le sens de discours que l'on a précédemment donné : discours comme espace spécifique d'organisation textuelle.

[2] Suzanne Bertrand-Gastaldi [1993, p. 147] note, sur ce point, que si « thésaurus et plan de classification sont tous deux le résultat d'une transposition d'énoncés antérieurs ou synchroniques, [le] corpus qui a servi à leur constitution est difficilement cernable, il est flou, non déclaré ».

[3] Notamment norme Z 47-100 (1981).

[4] Par exemple, Chaumier 1978.

[5] L'approche normative [norme Z 47-100 (1981)] ne définit pas explicitement ce qu'elle entend par « terme » ; dans certains cas, le terme est bien celui de la terminologie ; dans d'autres contextes, l'emploi paraît plus approximatif.

documents à indexer peut être « préféré » à un descripteur utilisé jusqu'alors : les deux termes peuvent être mis en relation d'équivalence.

Dans le cadre de l'expérience que nous avons réalisée[1], l'un des organismes documentaires a repéré un candidat-descripteur dans la source du *Monde* à indexer que nous avions retenue : il s'agit du terme « redressement judiciaire », employé à plusieurs reprises dans l'article consacré au verdict rendu par le Tribunal de commerce sur les entreprises dirigées par Bernard Tapie[2]. Le thésaurus utilisé par l'organisme documentaire comportait le terme « règlement judiciaire » ; ce terme est apparu aux indexeurs à la fois trop imprécis et trop en décalage avec le vocabulaire utilisé par les médias pour rendre compte de la longue série d'épisodes concernant le traitement judiciaire des sociétés de Bernard Tapie. Le terme « redressement judiciaire » a donc été constitué comme descripteur, le terme « règlement judiciaire » ayant été, lui, maintenu à ses côtés comme « synonyme ».

Dans certains cas, la mise à jour de thésaurus, par adjonction comme par suppression de descripteurs, peut amener à une réindexation de tous les documents indexés par les « anciens » descripteurs. On voit jusqu'où peut aller, dans les pratiques d'indexation, le brouillage de l'origine textuelle des descripteurs.

Sur ce point, nous nous attacherons à montrer que tout ce qui fait le poids « informatif » d'un terme tient précisément au discours auquel il renvoie : la différence introduite dans un thésaurus entre « règlement judiciaire » et « redressement judiciaire » n'est significative que si l'on connaît les emplois qui les distinguent, les dates, les lieux, les contextes politiques ou sociaux dans lesquels ils apparaissent.

Le rôle des discours dans la constitution d'un lexique documentaire, tout comme le statut originel des descripteurs comme mots de discours, sont complètement opacifiés par la notion de langage documentaire compris comme « langage artificiel ». Or il nous paraît essentiel de faire apparaître la dimension textuelle des descripteurs : c'est elle qui permet aux descripteurs de pouvoir fonctionner comme objets de discours ; c'est elle qui rend indispensable la réflexion sur la notion de corpus et de collection en indexation. C'est en ce sens que l'indexation pourra se comprendre comme une stratégie d'exploration des sources (§ II, ci-après).

B - Dimension énonciative en indexation

Tout comme la dimension textuelle, la dimension énonciative reste totalement implicite dans le discours classique.

Si, comme l'indiquent normes et manuels, les termes retenus dans la littérature pour figurer en tant que descripteurs doivent être déterminés par un « spécialiste du domaine », que deviennent ces termes lorsqu'ils sont extraits de leur contexte et utilisés dans le cadre de cercles larges de locuteurs ? Comment la « non-ambiguïté » qui les définit comme termes peut-elle encore rester active ?

Le discours normatif n'appréhende pas cet aspect de la question, tant il semble évident que la stabilité des termes acquise dans le cercle étroit des indexeurs au contact des textes ne peut que tout naturellement se maintenir jusqu'aux

[1] Voir Annexe 1.
[2] *Le Monde* du 1/12/1994, p. 24.

utilisateurs. Ni la question de l'utilisation documentaire des termes ni celle de leur mode de communication ne sont posées, parce que, nous semble-t-il, n'est pas posée la dimension énonciative de l'indexation.

En considérant l'indexation comme espace de discours constitué de ses propres objets (les documents) et de ses propres locuteurs (les indexeurs), on parvient à mettre au jour les enjeux de l'indexation : il s'agit moins de partager les mêmes mots que le même univers, le même espace d'utilisation des termes ; c'est par l'éclairage apporté par l'approche théorique de la pratique terminologique que l'on peut déterminer un tel enjeu.

Reste la question du devenir des terminologies utilisées dans des cadres où elles ne sont *a priori* plus valides. Cette question concerne tout autant l'approche théorique de la terminologie que l'approche théorique d'autres pratiques, qui, comme l'indexation ou la vulgarisation scientifique, utilisent des termes : « Si les termes apparaissent, sont formés et "institués" dans et par les discours spécialisés (scientifiques ou techniques), ils n'en figurent pas moins dans quantité d'autres discours, et c'est, en règle générale, leur fonctionnement dans la communication étendue à des non-spécialistes (ou des "moins spécialistes") qui pose avec le plus d'acuité les problèmes de leur interprétation, de leur fonction, de leur bien-fondé.[1] »

Le recours à l'évolution des problématiques en terminologie nous a permis d'approcher un aspect du discours documentaire en particulier : la dimension textuelle des descripteurs.

En terminologie, cette dimension apparaît comme ce qui constitue les termes dans leur usage et leur fonction professionnels : établir une communication spécialisée dans le cadre d'un cercle étroit de locuteurs. Quel rôle jouent, en indexation, les termes ? De quel type de communication relèvent les descripteurs ? Sur ce point, l'utilisation des terminologies dans les discours de vulgarisation scientifique apparaîtra particulièrement éclairante : elle permettra de préciser la dimension énonciative de l'indexation, que l'approche théorique de la terminologie avait permis de faire apparaître.

I.3 - Évolution des problématiques en vulgarisation scientifique : émergence de la notion de discours stratégique

Tout comme la pratique terminologique, la pratique du discours de vulgarisation a récemment bénéficié d'un effort de recherche, qui a fait apparaître une distinction entre le lexique utilisé (des termes « spécialisés ») et l'objectif de communication propre à ce type de discours.

L'enjeu, pour les théoriciens, est de montrer comment, dans le discours de vulgarisation, les termes peuvent rester des termes : c'est par la notion de « discours de reformulation » qu'est décrit le contact maintenu entre discours de vulgarisation et discours spécialisés (I.3.1). Comme dans le paragraphe précédent, nous nous essayerons à un rapprochement entre disciplines : le rapprochement proposé ici est plus inédit et moins évident que précédemment puisque, contrairement à la vulgarisation scientifique, l'indexation documentaire, proche en cela de la terminologie, ne produit pas de discours à proprement parler. Le rapprochement

[1] Mortureux 1995, p. 22.

nous semble néanmoins pouvoir être effectué sur la base des objectifs de communication, semblables, entre indexation et vulgarisation scientifique (I.3.2).

I.3.1 - PROBLÉMATIQUES DE L'ANALYSE DU DISCOURS DE VULGARISATION SCIENTIFIQUE

On essaiera ici de synthétiser l'essentiel des problématiques de l'analyse du discours de vulgarisation en mettant l'accent sur le déplacement qu'ont opéré les théoriciens[1] : l'analyse du discours de vulgarisation est passée d'une étude du lexique à une étude du discours[2]. Chemin faisant, les analystes se sont intéressés moins aux termes pour eux-mêmes qu'aux stratégies de communication dans lesquelles ils étaient utilisés.

En reprenant la définition qu'en propose Mortureux, on peut entendre par discours de vulgarisation : « Un ensemble d'énoncés qui ont en commun d'assurer la diffusion de connaissances en dehors des cercles étroits de spécialistes qui les produisent.[3] »

Les « vulgarisateurs », comme les appelle Jacobi, s'interrogent peu sur leur pratique[4] qui bénéficie d'une longue tradition[5]. Reste que leur pratique repose sur un « savoir spontané » qu'il peut être utile de mettre au jour notamment afin de constituer le discours de vulgarisation comme objet d'étude. C'est à cette entreprise que se consacre Jacobi [1987] dont nous ne retiendrons qu'un aspect de la critique de la perception des pratiques par elles-mêmes.

A - Le mythe du « troisième homme » : la traduisibilité de la science

Parmi l'ensemble des mythes ou postulats que Jacobi met au jour en examinant les représentations que les praticiens se font de leur profession, nous retiendrons celui qu'il nomme le « mythe du troisième homme », mythe qui repose à la fois sur une certaine conception de la science et sur une certaine conception de la langue : « les spécialistes emploieraient volontiers un jargon obscur ou incompréhensible par les tiers à seule fin de tenir hors de portée un savoir sur lequel se fondent leur compétence et leur autorité[6] ». Autrement dit, à aucun moment, la science n'est appréhendée par les praticiens sous l'angle de sa complexité conceptuelle (la seule complexité de la science est d'ordre linguistique) ; à aucun moment non plus, la langue n'est envisagée sous l'angle de sa complexité linguistique (la seule complexité de la langue vient des rapports de pouvoir ou de domination qu'elle instituerait par le biais des mots). On remarquera enfin que la complexité des

---

[1] Selon les chercheurs du domaine, il n'existe pas véritablement, du moins à ce jour, de « théorie de la vulgarisation », mais plutôt un « ensemble de travaux convergents qui définissent un champ », Jacobi 1988, p. 12.

[2] Conduite dans le cadre plus général de l'analyse de discours telle que Pêcheux a pu la définir (*supra*), Mortureux et Petit 1989 par exemple.

[3] Mortureux 1989, p. 43.

[4] Cf. Jacobi 1988, p. 13 : « La pratique vulgarisatrice ne fait pas, ou très peu, l'objet d'une distanciation de la part des vulgarisateurs eux-mêmes. C'est une pratique qui semble se suffire à elle-même sans autre justification que sa propre production. Dans l'ensemble, la pratique vulgarisatrice est une pratique qui ne se pense pas ».

[5] On s'accorde généralement à voir dans l'ouvrage de Fontenelle *Entretiens sur la pluralité des mondes* (1686) le premier exemple de texte de vulgarisation. Pour un aperçu historique, on peut se référer à Lazslo 1993, Jeanneret 1994 ou encore Raichvarg et Jacques 1991.

[6] Jacobi 1988, p. 20.

processus de communication dans lesquels la langue intervient n'est pas non plus abordée.

Sur la base de ce postulat d'une « traduisibilité de la science », le vulgarisateur se pose sous la figure du troisième homme, indispensable sauveur, entre homme de la science et homme de la rue[1]. Dès lors, l'essentiel de sa pratique repose sur une traduction lexicale d'un registre de langue à l'autre. C'est sur ce point qu'une analyse linguistique du discours de vulgarisation scientifique peut montrer que cette perception de la pratique vulgarisatrice est biaisée : « On a souvent comparé le travail du vulgarisateur à un travail de traducteur : il réécrirait la science, exposée par les chercheurs dans un jargon incompréhensible, avec les mots de tous les jours. Cette image, pourtant séduisante, est inexacte. Si la science aime à fabriquer des mots nouveaux, c'est que les mots usés et polysémiques de notre langue commune semblent comme incapables de dire la science. Et en remplaçant les termes scientifiques par des synonymes approximatifs, on ne peut que déformer, transformer, réduire, caricaturer, bref dénaturer la science.[2] »

Une analyse linguistique des discours de vulgarisation scientifique montre que, contrairement à toute attente, les vulgarisateurs recourent systématiquement, dans leurs discours, aux terminologies spécialisées. Cette présence récurrente des termes spécialisés passe souvent inaperçue, prise qu'elle est, dans le discours qu'ils construisent, par les stratégies de communication qu'ils déploient[3].

Dès lors que la vulgarisation scientifique ne se pense plus dans les termes d'une traduction mot à mot, son enjeu se laisse redéfinir : « Vulgariser est une entreprise qui se situe au cœur d'une contradiction : comme le scripteur se propose de faire connaître le sens des notions et des concepts spécialisés construits par les sciences, il est contraint d'utiliser les termes et lexies des langues de spécialité ; mais, en employant dans son texte des termes spécialisés, il redoute – à juste titre – que les lecteurs ne puissent en comprendre le sens ; pour prévenir les difficultés d'accès au sens des destinataires, le scripteur recourt à une série de mécanismes, de type métalinguistique le plus souvent, qui lui permettent de mettre en relation les termes scientifiques avec les mots connus de la langue commune.[4] »

B - Réintégrer la dimension discursive de la vulgarisation scientifique

Le courant de l'analyse linguistique[5] du discours de vulgarisation[6] s'est attaché à redéfinir le discours de vulgarisation scientifique dans des termes qui fassent apparaître le type de transformation que ce genre de discours opère : le discours de

---

[1] Jacobi 1988, p. 24 : « Ils [les vulgarisateurs] se perçoivent comme les intermédiaires naturels et essentiels chargés de combler l'écart entre les scientifiques et le grand public afin de rétablir une communication interrompue. Ils se présentent comme les artisans d'une possible révolution du savoir qui vise un partage plus équitable entre tous les acteurs sociaux ».

[2] Jacobi 1993, p. 81.

[3] Cf. *Ibid.*, p. 77 : « Cette fonction [reformulation discursive sur la base des termes] est à ce point intégrée au discours que ces opérations, à la différence du dictionnaire, qui les affiche ou les exhibe, sont comme cachées dans le texte ».

[4] *Ibid.*, p. 81.

[5] Nous n'abordons pas ici le courant sociologique de l'analyse du discours vulgarisateur qui examine les présupposés idéologiques d'une telle activité ; sur ce point, on peut, par exemple, se reporter à Boltanski et Maldidier [1977].

[6] Représenté entre autres par Mortureux [1983] et Jacobi [1984] pour les textes fondateurs.

vulgarisation a ainsi pu être défini comme une transformation discursive de textes sources[1].

L'analyste s'attache dès lors à examiner, dans le discours second (le discours de vulgarisation), les traces des discours premiers (les textes scientifiques)[2]. La présence des textes sources se signale par la présence de termes spécialisés dans le discours second. Comment identifier les termes puisqu'ils sont extraits du contexte de production qui les constitue comme termes ? Comment identifier les sources qui ne s'expriment, dans les discours seconds, que sous la forme des termes ?

Sans entrer dans le détail de la méthode d'analyse adoptée, on dira simplement que l'observation porte sur les stratégies discursives utilisées par les vulgarisateurs. Ces stratégies sont de deux types (définitionnel et désignationnel[3]), et se laissent identifier par une série d'indices de nature diverse, typographique (emploi des guillemets par exemple), linguistique (séquences paraphrastiques par exemple), etc. Les termes sont identifiés *via* le repérage de ces stratégies discursives menées sur un corpus de textes de vulgarisation. La méthode de l'analyse du discours de vulgarisation relève en effet de l'analyse interdiscursive, reposant sur l'hypothèse que le discours de vulgarisation scientifique s'inscrit dans un réseau de textes, défini par une « formation discursive[4] ». Cette analyse de l'emploi de termes spécialisés dans plusieurs discours de vulgarisation, comparé à l'emploi de ces mêmes termes dans plusieurs discours spécialisés, montre que le terme « scientifique » employé par les vulgarisateurs fonctionne différemment dans les textes scientifiques et dans les textes vulgarisés : dans les premiers, il révèle une certaine monoréférentialité (référence quasi constante au même concept) ; dans le second, il révèle une grande plasticité référentielle (variabilité des catégorisations référentielles).

Le rôle des termes en vulgarisation n'est donc pas comparable à celui qu'ils tiennent en terminologie : les termes, dans le texte de vulgarisation, sont là pour focaliser l'attention du lecteur sur les passages explicatifs ou définitoires, et non pour permettre au lecteur d'établir une relation référentielle stable[5]. Ils tiennent un rôle de « termes pivots » : comme isolés du discours par les stratégies discursives dont ils sont l'objet, ils apparaissent comme des citations, des extraits d'autres textes sur lesquels le texte présent s'appuierait pour définir, pour reformuler le savoir scientifique.

---

[1] « On sait que la diffusion des connaissances s'effectue notamment à travers une activité discursive de reformulation des sources dites "primaires" », Mortureux 1993, p. 3.

[2] Cf. Mortureux 1988, p. 119 : les discours de vulgarisation « sont des discours "seconds", dont la production, le fonctionnement et la légitimité renvoient à des « discours primaires » (dits parfois ésotériques), qui sont les publications par lesquelles les chercheurs exposent à leurs pairs les résultats de leurs travaux. Pour un linguiste, c'est là une première propriété de la vulgarisation scientifique, engageant à examiner, dans les énoncés réalisés, les modalités de cette secondarité, les manifestations de cette référence aux discours primaires, les traces de la réénonciation qui les a produits ».

[3] Nous y revenons au paragraphe II de ce chapitre.

[4] Jacobi 1988, p. 38-39. Pour une présentation de la notion empruntée à Foucault, voir le § II.1 de ce chapitre.

[5] Cf. Mortureux 1988, p. 145 : « La vulgarisation scientifique contemporaine ne fournit pas un apprentissage de la terminologie : les termes qu'elle cite sont généralement isolés des ensembles conceptuels et formels qui les structurent [...] et leur paraphrase n'attire guère l'attention sur cette systématicité des terminologies. [...] Le discours de vulgarisation scientifique se borne à faire référence, allusion, en mentionnant un certain nombre de termes. Là encore, on peut légitimement parler d'information plus que de formation ».

Ce que montre l'utilisation des termes dans les textes de vulgarisation, par opposition à l'utilisation des termes dans les discours spécialisés, c'est que le « savoir [y] est décontextualisé[1] ». Comme a pu également le souligner Grize, les sources scientifiques sont, dans le discours de vulgarisation, détournées de leur usage initial[2] ; et c'est grâce à ce détournement, à cette décontextualisation, que le « savoir » peut être transmis. En effet, par les déplacements référentiels opérés à partir des « termes pivots », le vulgarisateur déploie une série de désignations multiples qui permet des saisies et des ressaisies, diverses, des concepts scientifiques, par les lecteurs non spécialistes : ce n'est donc pas exactement le « même » savoir qui est diffusé entre spécialistes et à destination de non-spécialistes mais le lien est maintenu par le biais du terme, porteur d'un ancrage discursif dans les textes scientifiques.

Cette présentation de l'évolution des problématiques en vulgarisation scientifique, d'une approche centrée sur la traduction lexicale à une approche révélant des faits de discours et d'interdiscours, nous paraît de nature à préciser les enjeux de la dimension discursive en indexation.

I.3.2 - RAPPROCHEMENT DE DEUX DISCIPLINES

Le rapprochement entre indexation et vulgarisation scientifique, s'il semble être autorisé par la conception de la pratique vulgarisatrice que propose Jacobi[3], n'a pas été, à notre connaissance, établi.

On en trouve néanmoins, nous semble-t-il, une incitation dans une contribution que Sylvain Auroux présente dans un ouvrage collectif consacré aux sciences en bibliothèque[4]. Il émet l'hypothèse que l'accès au savoir dont la documentation a la charge doit s'inspirer des méthodes employées par la vulgarisation scientifique puisque l'enjeu consiste à éviter une « distorsion » entre les représentations de la science et les représentations communes : « ce dont nous avons besoin, c'est inventer et installer dans notre familiarité une forme de représentation admissible des connaissances qui soit en cohérence avec la pratique de notre vie quotidienne[5] ». Auroux illustre son propos ; le bibliothécaire apparaît comme celui qui doit établir une connexion entre un long article sur les particules et deux lignes sur ce sujet dans une encyclopédie : entre les deux, il y a, dit-il, « une chaîne de diffusion de savoir » à établir[6].

---

[1] Jacobi 1988, p. 22.
[2] Pour Grize, le propos de la vulgarisation consiste à « communiquer des savoirs à d'autres fins que leur mise en pratique », cité *in* Jacobi 1987, p. 8.
[3] Jacobi [1987, p. 31] tient que la vulgarisation scientifique n'est pas « une pratique spécifique » ; pour lui, « la vulgarisation scientifique s'inscrit dans un continuum de la diffusion de la science, elle en est une des modalités. [...] J'établis que la diffusion des connaissances peut [...] s'opérer par une diffusion structurée où un petit nombre de lecteurs deviennent des ré-utilisateurs d'une information soutenue. La diffusion large, auprès d'un public indifférencié, par le moyen de rhétorique particulière n'est probablement qu'un cliché dénué de consistance. Dans les faits, c'est une large panoplie de pratiques de socio-diffusion de la science qu'il faudrait évoquer », Jacobi 1987, respectivement p. 8 et p. 163.
[4] La contribution [Auroux 1994] a pour titre *Encyclopédies, bibliothèques et formalisation du savoir*.
[5] Auroux 1994, p. 150.
[6] *Ibid.*, p. 149.

Notons que le rapprochement de l'indexation et de la vulgarisation tend à redéfinir la finalité de l'indexation, voire plus largement, celle de la documentation : il s'agirait non plus de fournir des informations mais de donner accès aux savoirs.

La distinction entre information et connaissance constitue le cœur de débats importants dans le champ des sciences de l'information et de la communication[1]. L'angle d'approche que nous avons retenu ne permet pas réellement de donner corps à ces distinctions. En effet, nous avons jusqu'à présent essayé de préciser ce que, dans le cadre d'un modèle linguistique, pouvait être l'information telle que la norme d'indexation la présente : « extraite » d'un texte, « traduite » dans un langage documentaire, « trouvée » par un utilisateur. En dégageant cette formulation des présupposés réalistes dans lesquels elle semblait prise, nous avons proposé de comprendre l'information comme un « objet de discours » créé par l'utilisateur et l'indexation comme la pratique permettant de créer ces objets de discours. Quant à savoir si cet objet de discours permet à un utilisateur de s'informer ou de se former, nous ne disposons là d'aucun moyen de réponse.

En revanche, que l'on parle d'information ou de connaissance, la gestion des objets de discours en indexation pose des problèmes similaires. Il s'agit bien de maintenir une stabilité d'un bout à l'autre de la chaîne documentaire ; et c'est là que les remarques de Sylvain Auroux nous paraissent éclairantes.

En effet, qu'il s'agisse de représentations scientifiques ou de représentations « tout court », cet ensemble de représentations discursives que constituent les documents d'une bibliothèque ou d'un centre de documentation sont toutes singulières et ne correspondent pas *a priori* aux représentations, communes ou spécialisées, des utilisateurs, des lecteurs d'une bibliothèque ou d'un centre de documentation. En cela, la problématique de l'indexation rencontre celle de la vulgarisation : assurer une stabilité de la référence qui aille d'un auteur à un lecteur. Dès lors, l'indexation peut se penser sous l'angle de la vulgarisation scientifique sans que l'on ait forcément à se prononcer sur le caractère informatif ou pédagogique des accès qu'elle dispose. Son enjeu et sa finalité peuvent se laisser approcher sous l'angle du mode de diffusion des documents qu'elle réalise : c'est un mode de diffusion nécessairement élargi, nécessairement plus large que le contexte dans lequel les documents qu'elle propose ont été créés[2].

L'approche linguistique des discours de vulgarisation scientifique montre que le problème de la communication des savoirs vers des non-spécialistes n'est pas lié aux mots employés mais au contexte dans lequel ils sont utilisés et que tout l'effort du vulgarisateur porte précisément sur ce contexte. Autrement dit, l'approche linguistique de la vulgarisation fait émerger la dimension énonciative, complexe, de la communication vulgarisante et focalise l'attention sur les moyens linguistiques mis en œuvre pour opérer ce type de communication.

Il nous semble de même important de définir l'indexation en prenant en compte la diversité des foyers énonciatifs (relative aux auteurs comme aux indexeurs), diversité qui n'est pas réductible par l'adoption d'un système lexical réduit et

---

[1] Pour la période récente, Blanquet 1994 et Benoît 1992 proposent une synthèse des différentes positions sur les rapports entre information et connaissance en documentation.

[2] En ce sens, peut-être n'y a-t-il pas une grande différence entre la documentation, ou la bibliothèque, dite spécialisée et la documentation, ou la bibliothèque, dite encyclopédique : sans doute une nuance de degré peut-elle être établie (plus ou moins de domaines spécialisés en jeu, plus ou moins de types de lecteur différents).

unifié. Nous poserons donc, au cœur de la problématique du discours documentaire, cette pluralité d'univers de référence que doit permettre de traverser l'indexation. Dans les termes de Kripke, et comme nous le verrons ci-après, aussi bien en documentation qu'en vulgarisation, il s'agit d'opérer le passage d'*un monde possible*, où la fixité de la référence est acquise, à une *pluralité de mondes possibles*, à travers lesquels la référence peut se trouver modifiée.

Face à cette problématique commune à l'indexation et à la vulgarisation scientifique, quels sont les moyens que peuvent mettre en œuvre l'une et l'autre des deux pratiques ?

Comme nous l'avons vu, dans le discours de vulgarisation scientifique, le lien est maintenu avec les sources scientifiques initiales par le biais des termes, qui cependant ne jouent plus le même rôle que celui qu'ils tiennent dans les discours spécialisés : ils fonctionnent comme des « termes pivots », dotés d'une double fonction référentielle. Ils réfèrent d'une part aux sources initiales dans lesquelles ils sont définis proprement comme termes. Ils réfèrent d'autre part aux discours dans lesquels ils apparaissent et fonctionnent alors comme indicateurs du début d'une explicitation scientifique qui va procéder par accumulation de procédés rhétoriques divers (schéma, exemple, métaphore, etc.)[1].

Nous posons comme hypothèse que l'indexation consiste, elle aussi, à repérer, dans un document, des types de mot qui tiennent moins un rôle sémantique qu'un rôle désignatif, et un rôle désignatif double : les descripteurs référeraient, dans ce cadre, à la fois aux autres discours et au discours d'où il est extrait, en pointant sur les segments textuels aptes à permettre la construction de l'information recherchée. La formulation de cette hypothèse rejoint celle que nous émettions au chapitre II où nous proposions de penser le descripteur à travers la notion de thème de discours telle que l'approche Marandin. À ce stade de la recherche et compte tenu des acquis issus de l'analyse linguistique du discours de vulgarisation, nous pouvons étoffer cette hypothèse : le descripteur, s'il doit fonctionner comme un thème de discours, doit, compte tenu de la diversité des univers de référence qu'il doit permettre de traverser, présenter des caractéristiques proches de celles du terme de la terminologie : c'est sur ces bases que nous conduirons notre étude du descripteur dans le chapitre V.

Le rapprochement, par analogie, entre indexation et vulgarisation scientifique dessine le rapport que peuvent entretenir en indexation « langage documentaire » (entendu comme ensemble de descripteurs extraits d'une collection de documents) et « discours documentaire » (entendu comme ensemble de textes organisés selon des principes que nous définirons) : le langage documentaire est au discours documentaire ce que la terminologie est au discours de vulgarisation, une chaîne de traces lexicales destinée à la construction de l'interprétation.

Bien évidemment, le rapprochement entre indexation et vulgarisation, s'il peut paraître éclairant, se heurte rapidement à des limites : la vulgarisation scientifique peut se concevoir comme un discours second lié *via* les termes à des discours

---

[1] Par exemple, l'extrait suivant s'articule autour du terme-pivot « diapir » : « La cartographie structurale nous a révélé, il y a une dizaine d'année, l'existence, sous le Moho, d'une montée de manteau plus ou moins tubulaire. Par analogie avec les diapirs de sel, qui sont des colonnes irrégulières de sel montant spontanément grâce à leur densité inférieure à celle des sédiments qu'ils traversent, ces cheminées de matériel mantélique furent nommées "diapirs" », *La Recherche*, n° 239, janvier 1992, p. 25, cité *in* Jacobi 1994.

premiers parce qu'elle produit elle-même des discours ; l'indexation ne produit pas de discours dans ce sens-là. Il reste à définir en propre le discours documentaire.

Le détour analogique par l'examen de l'approche linguistique de la pratique terminologique et de la pratique vulgarisatrice nous a semblé utile pour deux raisons :

- *d'une part, il appuie la différence qui existe entre la perception de la pratique par elle-même et la perception de la pratique par des linguistes. La différence tient, dans les deux cas, à la focalisation extrême que les praticiens portent sur les mots au détriment des discours. En ce sens, il nous semble que les représentations que les indexeurs se font de leur pratique ne diffèrent pas des représentations d'autres praticiens ;*

- *d'autre part, il pointe ce qui, dans une pratique qui utilise des mots et des textes, reste souvent inaperçu : la dimension du discours sous ses deux aspects, textuel et énonciatif. Dans les deux cas, on voit que les termes entretiennent toujours des rapports étroits avec les textes dont ils sont issus. On voit aussi ce qui distingue mot et discours : ce ne sont pas les mots qui sont en terminologie « spécialisés » et en vulgarisation « didactiques » ; ce sont les discours dans lesquels ils sont employés qui sont spécialisés ou didactiques. De même, il importe de noter que ce ne sont pas les mots qui sont en indexation « informatifs » : c'est la façon dont ils se répondent de texte en texte qui peut créer de l'information, ou encore de la connaissance.*

Après les avoir perçues uniquement par le biais d'autres pratiques, nous proposons ci-après de préciser les problématiques propres au discours documentaire en indexation.

I.4- Approche du discours documentaire

Avant de définir le « cahier des charges » du discours documentaire, nous tâcherons de préciser l'enjeu de cette notion pour une approche de l'indexation : en quoi l'hypothèse de la création, par l'indexation, d'un espace de discours particulier, intermédiaire entre l'espace des auteurs et l'espace des utilisateurs, est-elle nécessaire ?

I.4.1 - ENJEU DE LA NOTION DE DISCOURS DOCUMENTAIRE

L'enjeu de la notion de discours documentaire, entendu comme espace d'organisation de documents, ne se pose que dans le cadre d'une certaine vision des institutions dans lesquelles l'indexation est pratiquée (bibliothèques et centres de documentation).

Sur ce point, on peut distinguer deux approches.

(i) La notion de discours documentaire est peu pertinente dans le cadre d'une approche qui considère la bibliothèque avant tout, et en vertu de son étymologie, comme un dépôt de livres, un lieu d'accumulation et de conservation, qu'il est nécessaire d'organiser simplement en raison du nombre des documents qui s'y trouvent. Les principes d'organisation sont donc dans ce cas établis *a posteriori* sur la base de critères qui peuvent être externes aux sources des documents ; ces principes sont le plus souvent « orientés

utilisateurs ». Il nous semble que relèvent de cette approche tous les discours où l'enjeu de la documentation et par suite de l'indexation se situe dans le traitement massif, toujours plus rapide, de documents toujours plus nombreux. Dans ce cadre, tous les moyens sont bons pour traiter les documents, les « indexer » : méthodes informatiques, statistiques, linguistiques, cognitives, ensemble ou séparées, ne sont évaluées qu'à l'aune d'un seul critère, la faillite ou le succès d'une recherche documentaire menée par un utilisateur : nous retrouvons ici l'expression de l'hypothèse de service que nous avions dégagée au chapitre I.

(ii) À l'opposé, et toujours à gros traits, se situe une approche qui considère la bibliothèque comme auxiliaire de la pensée humaine au sens large et vague du terme. Cette approche est diversement illustrée. On la précisera en reprenant la vision de deux épistémologues des sciences, Auroux et Latour, qui mettent l'accent sur la nécessité d'organiser les savoirs en bibliothèque non plus *a posteriori*, en rapprochant par des mots des documents déjà sélectionnés et constitués comme tels, mais *a priori*.

A - Auroux [1994] définit le rôle de la bibliothèque par rapport aux deux propriétés qui caractérisent, selon lui, le savoir scientifique : la formalisation et l'externalisation. C'est principalement sur la base de la seconde propriété d'externalité du savoir que la bibliothèque apparaît comme auxiliaire de la pensée : « le savoir a besoin pour fonctionner d'extension des capacités cognitives sous forme de bibliothèques, de discours de reformulation et de formalisation[1] ». Dans ce cadre, qui tient qu'il n'y a pas de savoir sans construction d'instruments externes, la bibliothèque constitue l'un de ces instruments. En ce sens, l'accès au savoir peut difficilement s'établir *a posteriori* sur la base d'une théorie, préexistante et externe aux sources, de la répartition des connaissances : « l'univers, soit réel soit intelligible, a une infinité de points de vue sous lesquels il peut être représenté et le nombre des systèmes possibles de la connaissance humaine est aussi grand que celui de ces points de vue[2] ».

On retrouve là une dialectique classique en documentation mais encore est-il nécessaire de la poser comme centrale, irréductible : si la bibliothèque est un instrument consubstantiel au savoir, son enjeu ne peut être celui de trouver un mode de représentation universelle d'un savoir, ce savoir n'existant pas sans, entre autres, la bibliothèque elle-même. Il s'agit plutôt de trouver des modes de représentation qui ménagent la diversité des points de vue qui, eux, permettront l'émergence du savoir.

Ce texte d'Auroux nous semble mettre en valeur la nécessité de penser la bibliothèque en termes d'espace spécifique et particulièrement en termes d'espace de représentations spécifique.

B - On retrouve chez Latour [1996], dans le cadre d'une problématique différente, la même expression de cette nécessité : « On comprend alors que les institutions comme les bibliothèques, les laboratoires, les collections ne soient pas de simples moyens dont on pourrait se dispenser aisément, sous prétexte que les phénomènes parleraient par eux-mêmes à la seule lumière de la raison. Additionnés les uns aux

[1] Auroux 1994, p. 147.
[2] *Ibid.* p. 145.

autres, ils composent les phénomènes qui n'ont d'existence que par cet étalement à travers la série des transformations.[1] »

L'argumentation que Latour défend contre la vision de la bibliothèque comme forteresse de papier ou comme empire de signes se déploie autour de la notion de « centre de calcul[2] ».

En qualifiant la bibliothèque de centre de calcul, Latour cherche à montrer l'articulation qu'opère la bibliothèque avec le « monde », avec son extérieur au sens large : elle est l'un des éléments qui participent à la construction de ce monde, à la construction de la réalité[3]. Là aussi elle se présente comme un auxiliaire, précieux, indispensable, de la pensée humaine. Son caractère indispensable tient à ce que, en tant que centre de calcul, la bibliothèque s'approche comme un lieu de transformation.

La transformation s'y réalise par un double mouvement de réduction et d'amplification[4] :

- la réduction est liée au fait que la bibliothèque ne manipule que des « inscriptions » dans les termes de Latour : par exemple, photo d'une situation, compte rendu d'une expérience scientifique et non situation elle-même ou expérience scientifique elle-même ; il s'agit, dans ces derniers cas, de « phénomènes » qui évoluent dans des lieux périphériques au lieu central du calcul ;

- l'amplification est liée au fait que la bibliothèque manipule des objets de même nature, susceptibles d'échanger entre eux des propriétés. Appréhendés sous une même forme, celle des inscriptions, les « phénomènes » au sens de Latour deviennent commensurables. Là où les phénomènes ne se donnent qu'en ordre dispersé, la bibliothèque donne la possibilité de les appréhender sous une forme unifiée : là encore il est question de regard, de « cohérence optique », dit Latour[5].

Dans l'approche de Latour, l'information se définit comme une relation entre les lieux périphériques (lieux des phénomènes) et le centre de calcul (lieu des inscriptions) ; en ce sens, l'information n'est pas un signe isolé, mais un signe chargé de la matière des phénomènes[6]. Elle est une inscription réduite par rapport à son lieu d'origine et amplifiée dans le nouveau lieu où elle se trouve : c'est par cette transformation que s'établit, selon Latour, la connaissance.

---

[1] Latour 1996, p. 41.
[2] La notion fait l'objet d'une description précise dans le cadre du modèle de Latour [1989] : on ne peut ici entrer dans les détails. Nous reprenons simplement l'approche qu'en propose Latour dans le texte qui nous occupe : « Pour comprendre un centre de calcul, il faut donc tenir du doigt l'ensemble du réseau des transformations qui relie chaque inscription au monde et qui relie ensuite chaque inscription à toutes celles qui lui sont devenues commensurables par la gravure, le dessin, le récit, le calcul ou, plus récemment, la numérisation », Latour 1996, p. 36.
[3] *Ibid.*, p. 28.
[4] *Ibid.*, p. 25-27.
[5] *Ibid.*, p. 33 : « Des informations éparses, provenant d'instruments épars peuvent s'unifier en une seule vision parce que les inscriptions possèdent toutes la même cohérence optique ».
[6] *Ibid.*, p. 25.

Ce qui permet l'émergence de la connaissance, c'est la « cohérence optique » à l'œuvre au niveau du centre de calcul. Cette cohérence optique n'est pas établie sur la base des signes proprement dits mais sur la base d'une compatibilité entre les inscriptions, qui permet aux inscriptions d'échanger entre elles des propriétés[1], de « capitaliser », dit Latour. En ce sens, l'inscription, ou l'information (les termes ne sont pas toujours distingués dans ce texte de Latour), se situe entre les mots et les choses ou plutôt traverse les deux mondes des mots et des choses : « Ces inscriptions circulent dans les deux sens, seul moyen d'assurer la fidélité, la fiabilité, la vérité entre le représenté et le représentant. Comme elles doivent à la fois permettre la mobilité des rapports et l'immuabilité de ce qu'elles transportent, je les appelle des "mobiles immuables" afin de bien les distinguer des signes. En effet, lorsqu'on les suit, on se met à traverser la *distinction usuelle entre mots et choses*, on ne voyage plus dans le monde, mais aussi dans les matières différentes de l'expression.[2] »

L'image de la bibliothèque que construit Latour est celle d'un laboratoire de transformation. Il compare d'ailleurs la bibliothèque à un laboratoire du CERN où sont « redistribuées les propriétés des phénomènes qui n'existent nulle part ailleurs et que savent saisir, repérer, amplifier des détecteurs géants construits pour l'occasion[3] ».

De ce texte foisonnant d'images, dans tous les sens du terme, nous retiendrons que la bibliothèque peut s'appréhender en termes d'espace de création de la réalité, ou plutôt d'une réalité. Elle est le lieu d'une rencontre et d'un échange entre phénomènes et inscriptions, entre les mots et les choses, un lieu où, par le format unique des inscriptions, la diversité des phénomènes a quelque chance de se trouver unifiée, pour peu qu'une « cohérence optique », qu'un « regard » puisse les faire voir comme semblables.

Dans ce cadre, l'indexation se conçoit comme ce « regard » qui permet aux phénomènes rapportés par les inscriptions de trouver une cohérence optique. Sur ce point, Latour souligne la nécessité de maintenir une distinction entre monde des signes, monde des inscriptions et monde des phénomènes, puisque seule cette distinction peut permettre de penser leurs interrelations[4]. Nous retrouvons en outre chez Latour une autre formulation de notre hypothèse concernant l'approche de l'information en termes d'objet de discours : sa notion de « mobile immuable » souligne la nécessité de penser l'information sous la double face d'une permanence et d'une instabilité référentielle.

Si nous avons largement mobilisé, depuis le début de ce chapitre, des auteurs issus de diverses disciplines, qui ne présentent pas toujours un lien direct avec notre sujet, si nous avons tenté de faire croiser ces différents regards, en prenant les risques de la démarche analogique, c'est que la dimension discursive de l'indexation, l'organisation spécifique des discours qu'elle réalise, n'est guère présente dans les approches classiques de l'indexation : elle reste entièrement à définir.

---

[1] Latour 1996, p. 38.
[2] *Id.*
[3] *Ibid.*, p. 39.
[4] *Ibid.*, p. 41 : « La vérédiction ne vient pas de la superposition d'un énoncé et d'un état du monde, mais provient plutôt du maintien continu des réseaux, des centres et des mobiles immuables qui y circulent ».

Pour contribuer à la définition de cet espace de discours documentaire, nous devions d'abord cerner ses différentes facettes et tenter de dégager son enjeu. Nous nous proposons désormais d'établir des référentiels (des cadres d'analyse, modèles ou théories) qui permettent de travailler ces différentes problématiques : nous les synthétisons au sein d'un « cahier des charges » du discours documentaire qui nous permettra de cerner le type de cadres théoriques nécessaires à retenir pour poursuivre notre recherche.

I.4.2 - LE DISCOURS DOCUMENTAIRE : « CAHIER DES CHARGES »

Nous avons pris comme point de départ la notion de langage documentaire, non pas tel ou tel langage documentaire (classification, thésaurus, liste de vedettes-matières, etc.), mais plutôt son principe : la représentation d'un texte établie au travers de mots considérés en eux-mêmes.

C'est d'abord dans la notion même de langage documentaire que la dimension discursive de l'indexation nous est apparue. En effet, la notion de langage documentaire, confuse en ce qu'elle ne distingue pas ses éléments, rend opaque les deux dimensions qui la constituent comme discours : la dimension textuelle et la dimension énonciative. Pourtant, la dimension textuelle est présente au tout début de l'indexation : les descripteurs qu'elle utilise sont toujours et d'abord des mots issus de discours, des mots du discours. De même, la dimension énonciative de l'indexation est-elle inscrite dans l'espace documentaire proposé à l'utilisateur. C'est en considérant les indexeurs comme formant un cercle étroit de locuteurs, dans lequel les mots peuvent trouver une certaine stabilité désignative, qu'apparaît l'enjeu de l'élargissement de ce cercle en indexation ; sur ce point, la problématique de l'indexation rejoint celle de la vulgarisation.

Penser l'indexation par le seul biais du langage documentaire paraît alors fortement réducteur : on n'y voit ni la dimension des textes ni celle des différents acteurs. Il nous semble qu'il faut alors observer l'amont du langage documentaire comme l'aval :

- en amont : c'est la sélection des sources elle-même qui doit retenir l'attention, car les descripteurs sont des mots issus de ces sources-là, et non de sources antérieures, passées, oubliées, jugées, le temps de constituer un langage documentaire, représentatives de toutes les sources à venir ;

- en aval : c'est l'organisation des documents, la façon de les soumettre au « regard », pour les rendre commensurables, pour permettre entre eux une circulation, qui mérite d'être observée ; là encore les mots ne sont pas seuls en jeu.

Sélection des sources et organisation des documents engagent une stratégie qui porte sur les discours eux-mêmes. Cette stratégie est, selon nous, double : il y a une stratégie qui concerne les sources (quels textes choisir ?) et une stratégie qui concerne les documents (comment montrer les documents ?).

Il y a plusieurs cadres d'analyse pour travailler ces deux aspects de la stratégie documentaire. Nous en avons retenu deux, sur la base de leur compatibilité avec une approche linguistique du lexique et de la référence, les deux problématiques qui nous paraissent centrales en indexation. C'est ainsi que pour donner quelques fondements à l'indexation comme stratégie d'exploration de sources, nous nous référerons au modèle des formations discursives de Foucault (II) ; pour donner

quelques fondements à l'indexation comme stratégie d'exposition des documents, nous emprunterons des éléments à la théorie des mondes possibles établie par Kripke (III).

II - Stratégie d'exploration des sources en indexation

Si l'indexation peut se penser sous l'angle d'un discours, entendu comme « ensemble d'énoncés et/ou de textes, possédant une organisation thématique, normative, structurale[1] », la question de l'indexation comme discours pose d'abord celle du type d'organisation qu'elle établit à partir de textes ou d'énoncés qui ne sont pas donnés *a priori* comme semblables, mais qu'elle donne à voir comme un « ensemble » (la collection documentaire).

Le discours documentaire est alors la forme du regroupement que réalise l'indexation à partir de sources hétérogènes. Cette forme de regroupement trouve généralement deux modes d'inscription : d'une part, sur un plan textuel, dans le document (regroupement, éclatement ou maintien des sources originelles[2]), d'autre part, sur un plan lexical, dans les descripteurs (qui matérialisent alors les liens entre documents).

Insistons sur le fait que ces modes d'inscription ne sont que des conséquences de l'opération de transformation discursive réalisée par l'indexation en amont, les conséquences du passage d'un type de discours (propre à une source, produite dans un certain contexte, promise à un certain usage) à un autre type de discours (propre au document, inséré dans un autre contexte, destiné à d'autres usages). Pour reprendre un exemple déjà cité, l'indexation est appréhendée ici comme ce qui fait passer un texte de la catégorie discursive « brevet » à la catégorie discursive « veille informative » quand ce texte est donné à lire comme porteur d'informations sur des innovations, et non plus comme attestation juridique protégeant un inventeur.

Il y a, schématiquement, deux façons de sélectionner et de regrouper des textes : soit *a posteriori* (c'est par exemple la notion classique de regroupement thématique), soit *a priori* (c'est par exemple la notion de formation discursive chez Foucault). Nous avons dégagé précédemment la nécessité, en indexation, de disposer d'un mode d'organisation qui soit établi *a priori*, sur d'autres critères que celui du « contenu » lui-même d'une source : paradoxalement, c'est là le seul moyen de ménager des points de vue différents sur des documents, tout en ayant la possibilité de déterminer des contours.

C'est en ce sens que la notion de « formation discursive » proposée par Foucault nous paraît nécessaire à disposer au centre de cette recherche comme « horizon théorique ».

---

[1] Souchard 1995, p. 258.
[2] Voir précédemment le compte rendu d'expérience, chapitre III § I.3., ou encore l'annexe 2 pour le détail.

La notion de formation discursive pourra paraître bien trop puissante pour couvrir les aspects très triviaux des problèmes documentaires. Il faudrait que les documentalistes et les bibliothécaires se livrent à un véritable travail scientifique sur les collections qui, s'il s'est fait pendant longtemps et continue à se pratiquer au sein de certaines bibliothèques[1], n'est pas toujours envisageable dans toutes les structures. Reste que le problème de la sélection et du regroupement, le problème plus général de la constitution de corpus s'avère, comme nous le montrerons, si déterminant pour l'attribution ou l'extraction *in fine* des descripteurs, qu'il semble nécessaire de poser au moins un cadre théorique à des modes de résolution qu'il restera à définir.

C'est dans cette optique que nous présentons dans ce paragraphe :

- d'une part, la notion de « système-archive » établie par Foucault, élément de son modèle des formations discursives qui nous paraît particulièrement adapté à une approche du discours documentaire à la recherche de fondements théoriques ;

- d'autre part, les modes d'exploration des sources habituellement utilisés par les professionnels et les problèmes qu'ils posent au regard de la notion de système-archive.

Le recours au modèle de Foucault pour examiner la notion de discours documentaire nous a été inspiré, d'une part, par les travaux réalisés en analyse de discours[2], d'autre part, par ceux réalisés en indexation par Suzanne Bertrand-Gastaldi[3]. Le modèle de Foucault a en outre été retenu dans la mesure où il nous paraissait compatible avec l'approche linguistique du lexique et de la référence qui nous sert de « toile de fond » théorique ; en effet, on lit par exemple dans Foucault [1966] : « Je voudrais montrer que le discours n'est pas une mince surface de contact, ou d'affrontement entre une réalité et une langue, l'intrication d'un lexique et d'une expérience ; je voudrais montrer sur des exemples précis, qu'en analysant les discours eux-mêmes, on voit se desserrer l'étreinte apparemment si forte des mots et des choses, et se dégager un ensemble de règles propres à la pratique discursive. Ces règles définissent non point l'existence muette d'une réalité, non point l'usage canonique d'un vocabulaire, mais le régime des objets.[4] »

C'est dans ce cadre que Foucault se propose de conduire une histoire des objets discursifs.

II.1 - La notion de « système-archive » dans le modèle de Foucault

Nous présenterons la notion de « système-archive » après avoir rappelé le projet de Foucault sous les aspects qui nous paraissent pertinents pour l'approche de cette notion dans notre recherche. Nous dégagerons, en fin de paragraphe, l'enjeu de la notion de « système-archive » en indexation.

---

[1] Notamment à la Bibliothèque nationale de France, qui développe un axe important de recherche sur les collections spécialisées ; un tel travail scientifique se mène aussi dans la plupart des bibliothèques dites d'étude.
[2] Marandin 1984 par exemple, mais aussi Jacobi 1988 par exemple.
[3] Bertrand-Gastaldi 1989 et 1993, par exemple.
[4] Foucault 1966, p. 66.

II.1.1 - PROJET DE FOUCAULT

Loin d'être spécialiste de Foucault[1], nous ne reprendrons sommairement que quelques éléments qui, de la problématique ou de la méthode, nous paraissent importants pour situer l'utilisation que nous ferons des concepts qu'il a établis.

A - Problématique

L'une des problématiques de Foucault peut se trouver résumée dans la question suivante : pourquoi ne peut-on pas dire que Darwin parle de la même chose que Diderot[2] ? Ou, sous une forme positive, au nom de quoi relie-t-on des auteurs qui ne se connaissent pas « dans une trame dont ils ne sont pas maîtres ?[3] ». Foucault, qui pose la question de la légitimité d'un corpus, s'attache alors à montrer l'insuffisance de la notion classique de « thème[4] ». Qu'il s'agisse de son travail sur la folie, le pouvoir ou la répression, le principe reste toujours de ne pas considérer d'emblée ces « mots » comme constituant des invariants, sur la base desquels des textes pourraient être rapprochés. Il s'agit non de dégager des permanences mais d'isoler les « lois » de coexistence des textes maintenus dans leur singularité.

La base de travail reste fournie par les regroupements thématiques habituellement effectués, mais l'enjeu consiste à s'en défaire, à déconstruire ces regroupements : « Certes, je prendrai pour repère initial des unités toutes données (comme la psychopathologie, ou la médecine, ou la philosophie politique) ; mais je ne me placerai pas à l'intérieur de ces unités douteuses pour en étudier la configuration interne ou les secrètes contradictions. Je ne m'appuierai sur elles que le temps de me demander quelles unités elles forment ; de quel droit elles peuvent revendiquer un domaine qui les spécifie dans l'espace et une continuité qui les individualise dans le temps ; selon quelles lois elles se forment ; sur fond de quels événements elles se découpent ; et si finalement elles ne sont pas, dans leur individualité acceptée et quasi institutionnelles, l'effet de surfaces d'unités plus consistantes.[5] »

Le matériau de base est donc constitué des « choses dites », dont il s'agit de dégager les règles de formation.

La notion de règles de formation des discours s'articule sur celle d'*a priori historique* : « l'histoire des choses effectivement dites jouit d'un *a priori* (historique) auquel l'archéologie assigne le rôle de rendre compte des énoncés dispersés[6] ». L'*a priori historique* qui donne la possibilité d'identifier des règles de formations discursives est un *a priori* qui ne concerne ni le sens ni la référence d'un énoncé mais plutôt sa « condition de réalité » (sa « positivité », dit encore Foucault) ; cet *a priori* est historique car il se donne dans le cadre d'une histoire spécifique.

---

[1] Nous nous appuierons dans ce paragraphe essentiellement sur Foucault 1969 et sur les lectures qu'en ont faites Marietti 1985 [1974] et Marandin 1984, 1993, 1997.
[2] Foucault 1969, p. 166.
[3] *Ibid.*, p. 167.
[4] « On aurait tort de chercher dans l'existence des thèmes les principes d'individuation d'un discours ». *Ibid.*, p. 51.
[5] *Ibid.*, p. 142.
[6] Marietti 1985 [1974] p. 155.

La notion d'*a priori historique* peut aussi être comprise, et c'est ce sens que privilégie Marandin[1], comme un « préconstruit », ce préconstruit pouvant valoir d'hypothèse pour la constitution d'un corpus[2].

Il ne relève pas d'une théorie de la langue mais d'« une théorie du fonctionnement de la langue dans une formation sociale[3] ». En cela, la notion de préconstruit se distingue de celle du « déjà-dit » qui porte, elle, sur le contenu. C'est ainsi que, dans le modèle de l'analyse de discours proposé par Pêcheux, l'interdiscours se comprend aussi en termes de préconstruit[4].

Reste que, dans les corpus établis sur la base d'un *a priori historique* ou encore sur la base d'un préconstruit, des *effets* d'invariance, comme des effets de déjà-dits, peuvent se rencontrer ; mais il s'agit là uniquement d'effets qu'on ne saurait prendre pour des principes de regroupement : « Cette forme de positivité [l'*a priori* historique] (et les conditions d'exercice de la fonction énonciative) définit un champ où peuvent éventuellement se déployer des identités formelles, des continuités thématiques, des translations de concepts, des jeux polémiques.[5] »

B - Méthode

Pour mener son projet d'identification des règles de formations discursives, Foucault se dote de trois méthodes d'analyse[6] ; seule l'une d'entre elle sera ici évoquée.

La notion de « pratique discursive » constitue une méthode d'analyse que Foucault a établie pour « suivre la formation des savoirs[7] ». Les discours sont, dans ce cadre, appréhendés comme des pratiques discursives, obéissant à ce titre à des lois de formation. Ces lois sont dégagées sur la base d'un examen des stratégies propres à une pratique. Par stratégie, Foucault entend aussi bien les « thèmes » spécifiques à une pratique que les « théories » élaborées au sein d'une pratique[8]. Thèmes et théories ne sont pas considérés comme points de départ mais comme points d'arrivée de l'analyse : l'analyse des pratiques discursives doit faire émerger un « niveau préconceptuel » par rapport aux discours eux-mêmes. La méthode d'analyse vise donc bien à capter le régime d'existence des objets de discours[9].

Thèmes et théories propres à une pratique sont alors extraits de leur domaine de validité courant, ils sont décontextualisés et reconsidérés dans un cadre nouveau : « Ce qui a été modifié [par l'enquête archéologique], c'est le rapport de ces affirmations [stratégies au sens de Foucault] à d'autres propositions, ce sont leurs

---

[1] Marandin 1984, 1993, 1997.
[2] Marandin 1984 oppose un jugement de ressemblance entre textes qui se ferait au niveau des « formes des énoncés » à un jugement de ressemblance qui s'élaborerait au niveau d'un préconstruit (appelé alors « pré-discursif »)
[3] Marandin 1993, [p. 10].
[4] Marandin 1997, p. 12.
[5] Foucault 1969, p. 167.
[6] Il s'agit de mener conjointement une analyse des pratiques discursives, une analyse des relations de pouvoir et une analyse des modes de reconnaissance des sujets.
[7] Foucault 1984, p. 12 : « Un déplacement théorique m'avait paru nécessaire pour analyser ce qui était souvent désigné comme le progrès des connaissances : il m'avait conduit à m'interroger sur *les formes de pratiques discursives qui articulaient le savoir* ». C'est nous qui soulignons.
[8] Marietti 1985 [1974] p. 41.
[9] *Ibid.*, p. 37.

conditions d'utilisation et de réinvestissement, c'est le champs d'expérience, de vérifications possibles, de problèmes à résoudre auquel on peut les référer.[1] »

Ainsi, par exemple, l'objet « folie » se constitue-t-il au sein de pratiques diverses touchant l'hospitalisation, l'internement, l'exclusion sociale, la jurisprudence, les normes du travail industriel et de la morale, etc.[2]. Il se constitue au gré des transformations que lui impriment ces pratiques diverses : « D'une façon générale, définir un ensemble d'énoncés dans ce qu'il a d'individuel consisterait à décrire la dispersion de ces objets ; saisir tous les interstices qui les séparent, mesurer les distances qui règnent entre eux – en d'autres termes, formuler leur loi de répartition.[3] »

La méthode d'analyse des pratiques discursives procède donc en deux phases : une phase de reconstitution d'une épistémologie (« la somme de ce que l'on a cru vrai ») et une phase de reconstitution du savoir (« ce dont on a pu effectivement parler »). Dans ce cadre, le savoir ne réside pas seulement dans les propositions scientifiques : il y a élargissement nécessaire à d'autres champs de discours[4]. S'il s'agit de s'attacher à ce qui fonde les thèmes ou les théories propres à une pratique, l'enquête archéologique ne consiste pas pour autant à restituer l'acte de fondation de ces thèmes ou théories[5].

L'analyse des discours en tant que pratiques discursives se situe donc à un niveau particulier, qui n'est ni le stade de l'avant-verbalisation d'une théorie ni celui de la verbalisation elle-même. Ce niveau que l'analyse reconstitue est celui qui permet de capter le moment où l'objet de discours apparaît (sous la forme d'un événement) et se fixe (sous la forme d'une chose) : c'est ce niveau propre à une pratique spécifique que Foucault nomme un « système-archive[6] ».

II.1.2 - NOTION FOUCALDIENNE DE FORMATIONS DISCURSIVES

La notion de « formation discursive », intimement liée à celle d'« archéologie du savoir » et de « positivité des discours », se laisse aborder de multiples façons[7].

---

[1] Foucault 1969, p. 136.
[2] *Ibid.*, p. 234 : « Cette pratique discursive, elle était investie dans la médecine certes, mais tout autant dans les règlements administratifs, dans des textes littéraires ou philosophiques, dans la casuistique, dans les théories ou les projets de travail obligatoire ou d'assistance aux pauvres ».
[3] *Ibid.*, p. 46-47.
[4] Marietti 1985 [1974], p. 123 : « Il faut comprendre que le savoir ne réside pas seulement dans les propositions scientifiques : l'enquête archéologique trouve son territoire aussi bien dans les textes littéraires ou philosophiques, dans les fictions, les réflexions, les récits, les règlements institutionnels, les décisions politiques ».
[5] Foucault 1969, p. 235 : « Les formations discursives, ce ne sont donc pas les sciences futures dans le moment où [elles sont] inconscientes d'elles-mêmes ».
[6] « Ce qui a pu être dit obéit à une loi que représente l'archive qui est le système qui régit l'apparition des énoncés comme événements singuliers. Définissant le système de l'énonciabilité de l'énoncé-événement, définissant le mode d'actualité de l'énoncé-chose, l'archive n'est autre que le système de son fonctionnement. Système général de la formation et la transformation des énoncés, l'archive fait surgir le niveau d'une pratique des énoncés dans leur valeur d'événements et de choses », Marietti 1985 [1974], p. 155.
[7] Ce n'est pas la moindre des difficultés que présente le texte de Foucault ; sur ce point, Marandin [1979, p. 48] relève : « Le texte de Foucault est diffus ; les notions sont définies par des suites de négations. Ce qui en rend la lecture difficile, et problématique toute tentative de description effective (sur des discours différents que ceux que Foucault a choisis et par quelqu'un qui n'est pas Foucault) ».

Dans la présentation qui suit, nous avons privilégié, en les simplifiant, trois approches : une approche en termes d'« espace », une approche en termes d'« archive » et une approche en termes de « domaine ». Sous ces trois angles se dégagent des moyens pour penser la notion de corpus, et, plus précisément, la problématique du regroupement des textes en indexation.

A - Formation discursive et création d'espace discursif

Une formation discursive permet de considérer un ensemble de discours comme formant un « espace limité de communication[1] ».

Considérée en termes d'espace discursif, une formation discursive présente les caractéristiques suivantes :

- c'est un espace polyphonique : devant « rendre compte des énoncés dans leur dispersion, dans toutes les failles ouvertes par leur non-cohérence, dans leur chevauchement et leur remplacement réciproque, dans leur simultanéité qui n'est pas unifiable, dans leur succession qui n'est pas déductible », l'espace discursif est, dans l'enquête archéologique, composé de sources textuelles d'origine nécessairement hétérogène : textes juridiques, littéraires, réglementaires, scientifiques, etc. De la même façon que le regroupement par thèmes avait paru douteux, le regroupement par genre ou par auteur est ici écarté[2] ;

- c'est un espace réglé : une formation discursive est déterminée par une même « loi de répartition », c'est-à-dire par un ensemble de règles qui décrivent, pour une période donnée, et dans une aire géographique donnée, les conditions d'existence des énoncés, c'est-à-dire leurs conditions de production et d'interprétation. Ces règles sont de nature exclusivement discursive : « s'il y a des choses dites – et celles-là seulement –, il ne faut pas en demander la raison immédiate aux choses qui s'y trouvent dites ou aux hommes qui les ont dites, mais au système de la discursivité, aux possibilités et aux impossibilités énonciatives qu'il ménage » ;

- c'est un espace historiquement réglé : la notion de formation discursive montre que le discours n'a pas seulement un sens et une vérité, il a aussi une histoire et une « histoire spécifique », propre à une « discipline[3] », ou mieux à un « domaine » (voir ci-après). En ce sens, le principe de la formation discursive relève d'un *a priori historique* qui est à distinguer des *a priori* formels : le premier ne peut rendre compte des seconds mais « permet de comprendre comment les *a priori* formels peuvent avoir dans l'histoire des points d'accrochage, des lieux d'insertion, [...] et de comprendre comment cette histoire peut être non point contingence [...] mais régularité spécifique ».

L'approche des formations discursives que propose Foucault à travers la notion d'espace limité de communication montre que l'exploration parmi les sources textuelles peut se faire sur d'autres bases que sur les notions de thème, genres

---

[1] Foucault 1969, p. 166.
[2] Sur l'absence de prise en compte de la notion d'auteur et d'œuvre chez Foucault, voir Marietti 1985 [1974], p. 99-120.
[3] Cependant, il n'y a pas de relation bi-univoque entre disciplines instituées et formations discursives, ne serait-ce que parce que l'archéologie du savoir se tient à distance de l'épistémologie, voir sur ce point Foucault 1969, p. 232-239.

textuels et auteurs. Il faut alors étudier les conditions de production historiques des écrits. L'approche en termes d'archive souligne elle qu'il s'agit là d'une *nécessité*.

B - Formation discursive et création d'un système-archive

La notion d'archive est définie ainsi par Foucault : « Au lieu de voir s'aligner, sur le grand livre mythique de l'histoire, des mots qui traduisent en caractères visibles des pensées constituées avant et ailleurs, on a, dans l'épaisseur des pratiques discursives, des systèmes qui instaurent les énoncés comme des *événements* (ayant leurs conditions et leur domaine d'apparition) et des *choses* (comportant leur possibilité et leur champ d'utilisation). Ce sont tous ces systèmes d'énoncés (événements pour une part, et choses pour une autre) que je propose d'appeler *archive*.[1] »

On retiendra notamment de cet extrait que le « système-archive » gère les énoncés sous leur double face d'« événement » et de « chose ». On pourra voir dans le type « événement » le contexte d'énonciation, c'est-à-dire les traces discursives que garde l'énoncé décontextualisé, et dans le type « chose », la possibilité pour un énoncé d'être autonome par rapport à son contexte de production, et prêt en cela à être « détourné » de son usage initial. Mais l'archive telle que Foucault la définit présente encore, dans le cadre de notre recherche, cet intérêt qu'elle correspond à un niveau particulier entre la langue et le corpus, un niveau « qui définit le mode d'actualité de l'énoncé-chose » : « Entre la *langue* qui définit le système de construction des phrases possibles, et le *corpus* qui recueille passivement les paroles prononcées, *l'archive* définit un niveau particulier : celui d'une pratique qui fait surgir une multitude d'énoncés comme autant d'événements réguliers, comme autant de choses offertes au traitement et à la manipulation.[2] »

Fondé sur des règles permettant « aux énoncés à la fois de subsister et de se modifier régulièrement », le système-archive se donne « comme un système général de la formation et de la transformation des énoncés ». Reste que l'archive d'une époque ne se capte jamais dans sa globalité : « elle se donne par fragments, régions et niveaux[3] ».

Par la notion de système-archive apparaît la possibilité de créer un niveau intermédiaire de discours où un texte apparaît à la fois en tant que « source » et comme « document », pour reprendre les termes que nous avions précédemment proposés. Il nous semble que l'indexation devrait s'attacher à créer un tel niveau qui donne à voir *la formation* des énoncés dans le cadre qui les constitue mais aussi *leur transformation*, quand ils sont rapprochés d'autres énoncés au sein de formations discursives.

Si l'indexation parvient à être la « transformation réglée de ce qui a été déjà écrit[4] », alors elle constitue des « domaines de savoir ».

---

[1] Foucault 1969, p. 169. (c'est nous qui soulignons).
[2] *Ibid.*, p. 171.
[3] *Id.*
[4] *Ibid.*, p. 183.

C - Formation discursive et création d'un domaine de savoir

La notion de « savoir » est définie par Foucault[1] sous les quatre dimensions suivantes :

- une dimension « sémantique » : un savoir, c'est « ce dont on peut parler dans une pratique discursive » ; c'est l'ensemble des objets discursifs manipulés par une pratique. Certains de ces objets pourront acquérir un statut scientifique mais un domaine de savoir ne saurait être réduit à ceux-là. Ainsi, pour reprendre l'exemple de Foucault, « le savoir de la psychiatrie, au XIX^{e} siècle ce n'est pas la somme de ce qu'on a cru vrai, c'est l'ensemble des conduites, des singularités, des déviations dont on peut parler dans le discours psychiatrique[2] » ;

- une dimension énonciative, celle des locuteurs : un savoir, c'est un « espace dans lequel un sujet peut prendre position pour parler des objets auxquels il a affaire dans son discours » ; en ce sens, la légitimité de la prise de parole dans un discours est élargie au-delà des seules autorités reconnues compétentes. Le sujet de discours n'est plus appréhendé comme un « auteur » propriétaire d'un savoir mais comme un « sujet situé et dépendant » ;

- une dimension intertextuelle, celle des relations entre textes : un savoir, c'est un « champ de coordination et de subordination des énoncés où les concepts apparaissent, se définissent, s'appliquent et se transforment ». En ce sens, et toujours en reprenant un exemple donné par Foucault, « le savoir de l'histoire naturelle, au XVIII^{e} siècle ce n'est pas la somme de ce qui a été dit, c'est l'ensemble des modes et des emplacements selon lesquels on peut intégrer au déjà dit tout énoncé nouveau » ;

- une dimension fonctionnelle, celle de l'usage : un savoir « se définit par des possibilités d'utilisation et d'appropriation offertes par le discours ». Ainsi, selon Foucault, le savoir de l'économie politique à l'époque classique « ce n'est pas la thèse des différentes thèses soutenues mais c'est l'ensemble de ses points d'articulation sur d'autres discours ou sur d'autres pratiques qui ne sont pas discursives[3] ».

En approchant la notion de formation discursive par le biais de la notion de domaine de savoir, Foucault souligne que le savoir ne peut se concevoir que dans le cadre de pratiques discursives données : « il n'y a pas de savoir sans une pratique discursive définie ; et toute pratique discursive peut se définir par le savoir qu'elle forme[4] ». De même il n'y a pas, non plus, d'objet de discours sans domaine de savoir, sans inscription dans une pratique discursive spécifiée.

En ce sens, si l'indexation, pour « fournir des informations », doit permettre de créer des objets de discours, ces objets de discours doivent nécessairement se laisser construire dans des domaines de savoir régis par des pratiques discursives spécifiques. En ce sens, un langage documentaire ne peut impunément remplacer un terme par un autre, sans que tout le pan de son inscription dans un discours ne

---

[1] Foucault 1969, p. 238.
[2] *Id.*
[3]. *Id.*
[4] *Ibid.*, p. 238-239.

disparaisse. Par l'approche du discours documentaire dans le cadre du modèle de Foucault, l'enjeu de l'indexation n'est définitivement plus du côté des mots.

Nous avons retenu de Foucault [1969] trois approches de la notion de formation des discours (approches en termes d'espace, d'archive et de domaine de savoir) qui devraient nous permettre de problématiser la notion de discours documentaire telle que nous la postulons.

II.1.3 - ENJEU DU MODÈLE EN INDEXATION

Foucault prend le soin de préciser que la notion d'archive qu'il établit n'a rien à voir avec la « somme de tous les textes » que pourrait conserver une bibliothèque : l'archive « ne constitue pas la bibliothèque sans temps ni lieu de toutes les bibliothèques[1] ».

Il y a, dans ces pages, une critique implicite de la conception classique des textes en bibliothèque, critique d'une conception naïve car purement réaliste qu'il nous a précédemment semblé important de dégager pour pouvoir s'en dégager. La bibliothèque ne considère le plus souvent les textes que comme les reflets des choses, des événements, sans pleinement prendre en compte l'épaisseur discursive des objets qu'elle détient : « [L'archive] c'est plutôt, c'est au contraire ce qui fait que tant de choses dites, par tant d'hommes depuis tant de millénaires, n'ont pas surgi par les seules lois de la pensée, ou d'après le seul jeu des circonstances, qu'elles ne sont pas simplement la signalisation, au niveau des performances verbales, de ce qui a pu se dérouler dans l'ordre de l'esprit ou dans l'ordre des choses ; mais qu'elles sont apparues grâce à tout un jeu de relations qui caractérisent en propre le niveau discursif.[2] »

Foucault attire notre attention sur la non-évidence à la fois de l'existence des textes et de leur coexistence. Que faire de cette mise en garde d'une part et de l'ancestrale ignorance de la spécificité discursive des objets en bibliothèque d'autre part ? D'un côté, il nous paraît délicat d'ignorer le rôle des pratiques discursives tel que Foucault l'a mis en valeur ; d'un autre côté, il paraît tout simplement impossible de conduire, en bibliothèque, une analyse des pratiques discursives « à la Foucault ».

Il nous paraît cependant possible de maintenir comme horizon théorique le système-archive établi par Foucault.

Le modèle proposé par Foucault, ici réduit à trois de ses dimensions, permet de dégager des éléments de nature à circonscrire les aspects du discours documentaire.

• *En tant qu'« espace limité de communication », polyphonique et historiquement réglé*, une formation discursive apparaît sous l'angle d'un « cadre interprétatif ».

Une formation discursive constitue une situation d'interprétation, un contexte pour un énoncé. Cette notion montre que l'on peut se passer des notions de thèmes, de genres textuels ou encore d'autorité d'auteurs pour relier entre eux des auteurs « qui ne se connaissent pas dans une trame dont ils ne sont pas maîtres ». Seul le maintien de la polyphonie, de l'hétérogénéité des textes, peut permettre

---

[1] Foucault 1969, p. 171.
[2] *Ibid.*, p. 170.

l'émergence de domaines de savoir. Dans ce cadre, le texte isolé apparaît comme une unité « faible[1] » : il doit être pris dans un ensemble.

La notion de règles discursives qui délimite un espace de discours peut se poser comme alternative théorique aux pratiques actuelles de la sélection documentaire prises dans une impasse, comme nous le détaillerons ci-après : la sélection des sources recourt habituellement aux langages documentaires. Or cette intervention des langages documentaires à toutes les étapes du traitement des documents (sélection, indexation proprement dite ou encore exposition des documents) empêche de considérer les sources sous leur angle discursif et rend biaisé, peu clair, le choix effectué parmi les sources : sont sélectionnées uniquement les sources dont peut rendre compte le langage documentaire.

• *En tant qu'« archive »*, la formation discursive se donne sous la forme d'un système établissant, pour un énoncé donné, une relation entre son contexte de production et ses potentialités d'utilisation (*cf.* la relation énoncé-événement et énoncé-chose), cette relation établissant un niveau intermédiaire entre langue et corpus.

La notion de système-archive permet de donner consistance à notre hypothèse d'un discours documentaire : le discours documentaire se définit alors comme un niveau discursif spécifique dont la particularité est de maintenir les textes sous leur double forme d'événement et de chose. À ce titre, le discours documentaire où pourraient se déployer les systèmes de formations propres aux pratiques discursives peut se concevoir comme une transformation discursive réglée.

• *En tant que domaine de savoir*, la formation discursive repose sur les quatre dimensions suivantes : sémantique, énonciative, intertextuelle et fonctionnelle.

Cette approche des formations discursives met l'accent sur des aspects du discours documentaire qui nous étaient précédemment apparus en ordre dispersé. Les objets de discours, que peuvent construire les utilisateurs sur la base de descripteurs et qui peuvent être alors des « informations » pour eux, nécessitent l'inscription dans un domaine de savoir : l'indexation doit pouvoir spécifier les auteurs des sources et des documents en termes de « sujet » du discours, mais aussi établir des relations d'intertextualité et permettre un détournement des sources. L'émergence du savoir se réalise en effet à la marge des domaines constitués, par empiétement sur des textes dont l'appréhension première ne laisse pas forcément penser des formes de rapprochement, qu'en revanche une pratique discursive peut être amenée à établir.

Appréhendée sous trois facettes, la notion de formation discursive proposée par Foucault donne corps à la notion de discours documentaire comme espace d'organisation spécifique des documents, en montrant que l'on ne peut impunément s'en remettre aux textes tels qu'ils se donnent. Le modèle de Foucault oblige à penser, à expliciter les bases sur lesquelles s'effectuent l'exploration et le regroupement des sources en indexation : c'est d'abord cette contrainte de l'explicitation des principes de sélection qu'il nous semble important de retenir. La pratique d'indexation peut difficilement, nous semble-t-il, faire l'impasse sur cette question. En ce sens, il apparaît que, parmi les fondements théoriques de

[1] Foucault 1969, p. 34 : « L'unité matérielle du volume n'est-elle pas une unité faible, accessoire, au regard de l'unité discursive à laquelle il donne support ? ».

l'indexation, doivent figurer des éléments touchant aux principes d'exploration des sources ; à ce titre, le modèle de Foucault fournit un cadre possible.

Par ailleurs, le modèle de Foucault insiste sur la nécessité de concevoir un énoncé sous l'angle de ses utilisations nouvelles. Il y a sur ce point aussi, dans le modèle de Foucault, des éléments pour un fondement théorique de l'indexation vue comme une opération de détournement d'usage des sources. Enfin, le modèle de Foucault nous permet de dégager une autre dimension susceptible, elle aussi, de participer aux fondements théoriques de l'indexation : c'est la nécessité de considérer les objets de l'indexation comme des objets discursifs, sous l'angle de leur épaisseur discursive.

Autrement dit, en posant dans l'horizon théorique de l'indexation la notion de système-archive ou de formation discursive au sens large, on est conduit à considérer l'indexation dans ses choix initiaux, qui sont d'ordre discursif, et non plus dans ses choix finaux, qui sont de nature lexicale : des documents ne peuvent être jugés semblables parce qu'affectés du même descripteur.

II.2 - Le « système-archive » comme horizon théorique

En posant la notion foucaldienne d'archive à l'horizon de notre réflexion sur les fondements théoriques de l'indexation apparaissent des caractéristiques de la pratique documentaire rarement explicitées dans les discours de la pratique sur elle-même. En effet, mené à la lumière du modèle de Foucault, l'examen des pratiques d'indexation révèle, sous une forme parfois très ténue, les marques d'une « transformation discursive ». Pour faire apparaître cet aspect de l'indexation, nous poserons sous forme de conjecture que le processus de l'indexation relève d'une « fonction énonciative » (II.2.1). Nous tâcherons ensuite de mettre au jour les types de règles à l'œuvre dans le travail d'indexation compris dans sa phase de sélection des sources (II.2.2).

II.2.1 - CONJECTURE : L'INDEXATION COMME « FONCTION ÉNONCIATIVE »

Le processus de l'indexation partage avec le « système-archive » de Foucault le même objectif : faire voir un texte sous sa double face d'événement singulier (source) et de chose manipulable (document).

Pour atteindre cet objectif, l'indexation réalise une transformation qui porte, non sur le texte lui-même, mais sur son « entourage », son contexte ; en ce sens, on peut la rapprocher de la notion de « fonction énonciative » proposée par Foucault, fonction qui permet à un texte d'exister comme un « énoncé ». Le document pourra être vu, en indexation, comme un énoncé dans ce cadre-là et l'indexation comme la fonction d'existence d'une source comme document.

La fonction énonciative est approchée par Foucault dans les termes suivants : « La fonction énonciative – montrant bien par là qu'elle n'est pas pure et simple construction d'éléments préalables – ne peut s'exercer sur une phrase ou une proposition à l'état libre. Il ne suffit pas de dire une phrase, il ne suffit même pas de la dire dans un rapport déterminé à un champ d'objets ou dans un rapport déterminé à un sujet, pour qu'il y ait énoncé – pour qu'il s'agisse d'un énoncé : il faut la mettre en rapport avec tout un champ adjacent. [...] Un énoncé a toujours des marges peuplées d'autres énoncés. [...] Il n'y a pas d'énoncé en général, d'énoncé libre, neutre et indépendant mais toujours un énoncé faisant partie d'une série ou d'un ensemble, jouant un rôle au milieu des autres, s'appuyant sur eux et se distinguant

d'eux, il s'intègre à un jeu énonciatif, où il a sa part aussi légère, aussi infime, qu'elle soit.[1] »

On peut observer, dans les pratiques d'indexation, les traces de l'exercice de l'indexation comme fonction énonciative :

(i) chaque pratique d'indexation établit une organisation spécifique des documents. Indexé par un même mot, un document n'est pas pour autant introduit dans la même « série » pour reprendre les termes de Foucault : le « champ adjacent » retenu pour un énoncé-document n'est pas toujours le même. Il y a là, nous semble-t-il, la marque d'une fonction énonciative : l'indexation se signale par la série des documents qu'elle constitue (II.2.1.1) ;

(ii) les pratiques d'indexation utilisent divers procédés pour maintenir l'aspect dual du texte (source et document) : on observe des moyens pour maintenir la trace des conditions de production d'une source ; on relève aussi les marques d'un « jeu énonciatif » qui permet à un document-énoncé de trouver sa place dans une série. C'est en ce sens que le document pourra apparaître comme un « énoncé » au sens foucaldien du terme (II.2.1.2).

En affaiblissant les concepts établis par Foucault, nous proposons ci-après une analyse des pratiques d'indexation qui permet, sinon de définir à proprement parler l'indexation comme une fonction énonciative, du moins d'exhiber des mécanismes discursifs, tenus le plus souvent hors du cadre d'appréhension de l'indexation, comme le fait remarquer Suzanne Bertrand-Gastaldi : « En somme, les "traductions successives" des textes scientifiques par l'homme et la machine font progressivement disparaître le discours au profit du vocabulaire, la syntaxe et les indices au profit des symboles. Alors que le discours est une actualisation de la langue, il est presque retransformé, dans les bases de données, en langue artificielle, reconstituable à partir des formations discursives. Il reste cependant des traces irréductibles des conditions d'énonciation.[2] »

Si nous rejoignons Suzanne Bertrand-Gastaldi sur la nécessité d'une reformulation de l'indexation dans des termes qui fassent voir sa dimension discursive, nous ne pensons pas que, en l'état actuel, les bases de données et les types d'accès qui s'y trouvent puissent permettre une reconstitution aisée des « formations discursives » : nous y reviendrons.

II.2.1.1 - L'organisation des discours en indexation

Si, comme nous l'avons précédemment mentionné, la littérature courante sur l'indexation relève régulièrement la variabilité des termes utilisés pour indexer les mêmes textes, elle relève moins la variabilité des textes qui se trouve sous l'indexation par un même terme.

Pour illustrer cet aspect, nous prendrons d'abord un exemple extrait de l'expérimentation que nous avons menée[3].

---

[1] Foucault 1969, respectivement p. 128 et p. 130.
[2] Bertrand-Gastaldi 1989, p. 22.
[3] L'expérimentation est présentée en annexe 1.

Le même descripteur, « Arte », est utilisé par six organismes documentaires pour indexer un même document issu du journal *Le Monde*[1]. Or, ce même descripteur ne renvoie pas toujours aux mêmes textes. On retrouve ici la variabilité de mise en documents que nous avions précédemment relevée, mais le point consiste ici à remarquer que cette variabilité n'est pas exprimée par le biais du descripteur : la variabilité des mises en discours reste invisible aux yeux d'un utilisateur.

Ainsi, si l'on imagine la requête suivante adressée aux six organismes documentaires analysés – « Dans *Le Monde* du 1er décembre 1994, qu'y a-t-il eu sur Arte ? » –, on obtiendra trois types de réponse différents :

(i) un article : pour quatre des six organismes documentaires, seul l'un des deux articles du *Monde* est affecté du terme « Arte » ;
(ii) deux articles : pour un centre de documentation, les deux articles sont chacun indexés par le terme « Arte » ;
(iii) quatre articles : l'un des six organismes documentaires a intégré les deux articles du *Monde* dans une revue de presse comportant des articles issus d'autres sources.

Encore ne mesurons-nous ici la variabilité de mise en discours uniquement en considérant une seule source d'information.

En interrogeant les catalogues de bibliothèque qui utilisent le même langage documentaire, le langage Rameau, on retrouve la même variabilité de mise en discours sous l'emploi d'un même terme. Ainsi, à l'interrogation par le terme « analyse documentaire », on trouve le même ouvrage (Chaumier 1988) mais alors qu'il se situe au sein d'une série de 22 éléments dans la bibliothèque de l'ENSSIB, il est inscrit dans une autre série de 6 éléments dans le catalogue du réseau BRISE à Saint-Étienne. Certes, les deux bibliothèques ont procédé à la même sélection de source qu'elles présentent par le même accès, mais l'inscription de l'ouvrage dans un ensemble est radicalement différente. Cette différence traduit une variation dans les principes de sélection des ouvrages retenus, variation qui n'est pas rendue visible aux utilisateurs alors qu'elle constitue, nous semble-t-il, la marque même de l'indexation, voire tout son intérêt. Une indexation doit permettre de fournir un regard problématisé sur la production éditoriale ; or si problématisation il y a, sous la forme de règles de sélection, elle est maintenue inaccessible.

Ces deux exemples montrent que l'indexation réalise une organisation des discours spécifique, qui présente la singularité de ne laisser aucune « trace » visible ; ce n'est pas pour autant qu'elle ne suit aucune règle. En examinant les éléments qui constituent le contexte d'un document, on peut relever des « traces » qui indiquent que la « fonction énonciative » réalisée en indexation, si elle n'est pas entièrement réglée, est du moins partiellement régie par les sources elles-mêmes.

II.2.1.2 - Traces de décontextualisation et de recontextualisation dans les objets documentaires

À la fin du chapitre III, nous avons proposé de considérer l'indexation comme une opération de contextualisation réalisée en deux temps : à une décontextualisation des sources succède une recontextualisation au terme de laquelle une source devient

---

[1] *Le Monde* du 1er décembre 1994 propose en page 10 deux articles : le premier s'intitule « ARTE veut élargir son public et casser son image de chaîne élitiste », le second, un encadré, porte le titre « Jean-Marie Cavada, premier président du GIE ».

un document. Cette première approche de l'indexation trouve un fondement théorique dans la notion de système-archive proposée par Foucault, qui établit précisément un niveau où les deux faces d'un texte – événement et chose – peuvent être captées : c'est, dans le modèle de Foucault, le niveau où se forment les objets discursifs propres à une pratique discursive.

S'il est clair que les pratiques d'indexation sont encore loin de constituer, dans leur espace de discours, ce niveau des formations discursives, il semble qu'elles mettent en œuvre des procédés pour maintenir un texte sous son double aspect de source et de document.

A - Marques de la source dans le document : traces de la décontextualisation

Dans la chaîne de traitement documentaire telle qu'elle est traditionnellement présentée dans les manuels ou les traités, l'opération d'indexation est distinguée de celle du catalogage (ou description bibliographique) qui consiste à prélever, de l'entourage d'une source, des données que l'on nomme factuelles et qui comprennent des mentions relatives aux titres, auteurs, éditions, dates, pagination, langue, etc.

Dans le cadre d'une approche de l'indexation en termes de transformation discursive, cette formalisation de données externes au « contenu » du texte participe à la création du document. Elle correspond, nous semble-t-il, à une tentative pour maintenir, dans le cadre du document, quelques-unes des particularités du contexte de production d'une source. Cette présence des éléments de la source au sein du document n'est pas sans incidence sur l'interprétation des termes d'indexation, comme le remarque Suzanne Bertrand-Gastaldi : « Deux indexations en tout point semblables, de par leur rattachement à deux documents différents publiés dans des revues différentes, à des dates différentes, constitueront bien deux énoncés différents qu'il faudra bien interpréter différemment avec l'aide du co-texte et de l'intertexte qui, lui, contient donc aussi un certain nombre d'éléments indiciels.[1] »

En ce sens, il nous semble que les pratiques documentaires parviennent, dans une certaine mesure, par le biais de la description bibliographique, à maintenir l'hétérogénéité des sources.

Sur ce point, il nous semble que le document doit être en indexation défini comme un objet complexe, comprenant, outre le texte de la source et les descripteurs, les données factuelles issues de la description bibliographique.

Appréhendé sous cet angle, un document peut être compris comme un énoncé, du moins dans le sens foucaldien du terme. En effet, selon Foucault, il n'est d'énoncé que situé, et situé doublement : par rapport à son contexte de production d'une part[2] et par rapport à son contexte d'utilisation d'autre part[3].

---

[1] Bertrand-Gastaldi 1989, p. 12.
[2] Foucault 1969, p. 13 : « Les coordonnées et le statut matériel de l'énoncé font partie de ses caractéristiques intrinsèques. [...] Il faut qu'un énoncé ait une substance, un support, un lieu et une date. Et quand ces conditions requises se modifient, il change lui-même d'identité ».
[3] *Ibid.*, p. 130 : « Il n'y a pas d'énoncé en général, d'énoncé libre, neutre et indépendant mais toujours un énoncé faisant partie d'une série ou d'un ensemble, jouant un rôle au milieu des autres, s'appuyant sur eux et se distinguant d'eux ».

En ce sens, le document, si on l'entend comme un texte pourvu de ses « coordonnées », peut fonctionner comme un énoncé puisqu'il permet d'établir ainsi un lien avec la source d'où il provient. En effet, comme le précise encore Foucault, « une série de signes deviendra un énoncé à condition qu'elle ait à "autre chose" un rapport spécifique qui la concerne elle-même, et non point la cause, non point ses éléments[1] ». Tel que nous l'avons envisagé jusqu'à présent, le document nous paraît être un énoncé dans la mesure où il entretient une relation avec « autre chose », ce que nous avons nommé la « source », cette relation ne relevant ni de la causalité ni du définitionnel, mais de l'interprétation : nous avons en effet posé le document comme étant l'interprétant de la source[2].

Ouvrons ici une parenthèse qui soulignera à nouveau le rôle opacifiant du langage documentaire sur la perception des sources en indexation.

Le document peut permettre l'interprétation d'une source grâce aux autres documents auquel il y est lié *via* les descripteurs. Sur ce point, les descripteurs sont considérés, dans l'approche de l'indexation que nous proposons, comme des marques de l'insertion d'un document dans une série plus que comme des marques de son contenu propre. L'approche classique des langages documentaires, et des thésaurus en particulier, tend à obscurcir ce phénomène. En effet, selon les définitions normatives, la particularité des thésaurus tient aux relations qui y sont établies entre les termes[3] (relations de généricité, de spécificité, d'équivalence ou d'association) : c'est la stabilité de ces relations qui est posée comme garante d'une « bonne » indexation. Or, comme le souligne Suzanne Bertrand-Gastaldi, les relations qu'entretiennent les descripteurs au sein d'une base de données ne sont plus les mêmes que celles proposées dans le thésaurus[4]. Rien de plus normal puisque le corpus qui a permis d'établir le thésaurus n'est pas le même que celui qui est soumis à l'indexation.

B - Marques du document : traces de la recontextualisation

La recontextualisation s'exprime prioritairement par le biais des descripteurs qui signalent l'inscription d'une source dans un nouvel ensemble où vont pouvoir se déployer de nouvelles utilisations, de nouvelles possibilités d'interprétation : « L'insertion d'un texte sous une rubrique de classification, d'un descripteur ou d'un ensemble de descripteurs par un indexeur donne un nouvel éclairage à ce texte, crée un nouveau réseau de relations en le rapprochant à la fois du système de classification ou du thésaurus et de l'ensemble des autres textes qui ont déjà reçu les mêmes "étiquettes".[5] »

Dans ce cadre, le nouveau contexte d'un document est constitué par l'ensemble des autres documents regroupés sous le même descripteur. C'est à ce stade que la

---

[1] Foucault 1969, p. 117.
[2] Voir précédemment chapitre III § III.1.
[3] Cf. norme Z 47-100 (1981), p. 7 : « Une des fonctions primordiales d'un thésaurus est de représenter les relations entre concepts par l'indication des rapports entre les termes utilisés pour les décrire. Le réseau des relations d'un descripteur avec les autres termes (descripteurs ou non-descripteurs) fournit ainsi une sorte de définition et concourt à réduire les risques d'ambiguïté en situant le descripteur dans un contexte qui en précise le sens ».
[4] Bertrand-Gastaldi 1989, p. 12 : « Chaque mot-clé attribué à un texte entretient avec les autres des relations syntagmatiques et l'ensemble des mots-clés attribués dans une banque de données n'a rien à voir avec l'ensemble des descripteurs d'un thésaurus et de leurs relations paradigmatiques ».
[5] Bertrand-Gastaldi 1993, p. 145.

source acquiert une existence spécifique, une « existence documentaire » (une existence en tant que document), qui nous paraît être le résultat spécifique de la fonction « indexation », sa spécificité tenant principalement dans le « détournement d'usage » qu'elle met en place. Il importe à nouveau de rappeler cette évidence qu'il n'existe aucune source explicitement dédiée à l'indexation et que seule une fonction de « transformation » peut permettre de créer un document.

Si la marque la plus visible de la « fonction énonciative » par laquelle nous nous proposons de décrire l'indexation reste les descripteurs, d'autres aspects se révèlent au niveau spécifique de l'espace documentaire. En effet, si la source est inscrite dans le document sous la forme de données factuelles, le document est aussi inscrit dans le choix des sources : soit sous la forme directe d'une relation d'intertextualité entre sources à sélectionner et documents déjà constitués, soit sous la forme indirecte du thésaurus, où le poids du corpus présidant à sa création influence le choix des sources retenues. Nous abordons ce dernier aspect ci-après dans le cadre de la problématique plus générale des principes d'exploration des sources en indexation.

Comme dans les paragraphes précédents, où l'on a mené une analyse des pratiques d'indexation sous l'angle de la notion de « fonction énonciative », c'est l'éclairage fourni par la notion de « système-archive » établie par Foucault qui nous permettra de problématiser les règles de sélection en indexation.

II.2.2 - PROBLÉMATIQUE DES RÈGLES D'EXPLORATION DES SOURCES EN INDEXATION

La question des règles orientant le choix des sources et légitimant leur existence en tant que documents, que ce soit en bibliothèque ou dans les centres de documentation, constitue une préoccupation récente dans le milieu professionnel, comme le note Bertrand Calenge[1], auteur d'un des rares ouvrages français sur les « politiques d'acquisition ». A l'heure actuelle, les structures documentaires pourvues de « chartes » explicitant les critères de choix restent beaucoup moins nombreuses que celles qui en sont dépourvues. Un véritable « tabou » pèse sur le processus d'acquisition, remarque Calenge[2].

L'enjeu est pourtant de taille. Les bibliothèques et les centres de documentation abondent d'anecdotes relatant le poids des affinités électives d'un individu pour un domaine ou un auteur, créant à terme des trous béants dans les collections documentaires comme des sur-représentations inexplicables de certains sujets ; de quoi rendre incohérent le résultat de requêtes documentaires, la meilleure indexation possible ne pouvant guère pallier le déséquilibre des collections elles-mêmes. De façon moins anecdotique, une enquête menée auprès des publics de la Bibliothèque nationale[3] a montré que la faible fréquentation des chercheurs dans le domaine dit des sciences « dures » est directement liée à une absence de collections les concernant[4].

---

[1] Calenge 1994.
[2] *Ibid.*, p. 24.
[3] Étude non publiée menée sous la direction de Christian Baudelot. Une synthèse se trouve dans Baudelot et Véry 1994.
[4] En effet, pendant près de trente ans, la Bibliothèque nationale a délaissé l'acquisition de documents en langue étrangère dans le domaine des sciences « dures ».

Ces dernières années, plusieurs facteurs ont contribué à la création de chartes destinées à expliciter les règles d'exploration des sources, notamment :

- les impératifs budgétaires qui frappent tous les établissements culturels ont définitivement mis fin à toute velléité d'exhaustivité ;

- la notion de réseau documentaire a permis aux structures documentaires de penser les acquisitions en termes de complémentarité, tandis qu'elle exigeait une définition claire des champs de connaissances investis et des critères de sélection ;

- la multiplicité des types de structures documentaires, et notamment l'émergence des bibliothèques de lecture publique, a modifié la perception des bibliothèques et de leurs fonds. Il ne s'agit plus d'accumuler des documents, dont la pertinence sera bien découverte, un jour, par un chercheur, mais de présenter une collection cohérente susceptible de fournir des orientations claires à ceux que la sociologie des publics en bibliothèque nomme les « faibles lecteurs » ;

- plus globalement, il est apparu aussi et essentiellement que la notion de collection documentaire et celle de politique d'acquisition qui la sous-tend constituaient en propre le métier des professionnels de l'information et de la documentation[1].

On dispose désormais, pour une part, de règles explicites d'exploration des sources : on synthétisera celles que propose Calenge. On présentera aussi celles qui nous ont été fournies au cours de notre enquête, qui concernent plus spécifiquement les critères qui, pour une source donnée, guident le choix des segments textuels à indexer.

Reste une part importante de règles implicites, que des chercheurs en sciences de l'information ont fait émerger : nous joindrons à leurs remarques celles que nous avons pu faire au cours de notre expérience.

II.2.2.1 - Les règles explicites d'exploration des sources

Qu'il s'agisse de sélection de sources (A) ou de sélection d'objets textuels à indexer (B), c'est le recours à un référentiel existant qui est le plus souvent utilisé par les professionnels. Ces référentiels sont constitués des grands domaines de connaissance ou d'une liste de thèmes jugés spécifiques au public d'un établissement. C'est sur ce point que l'on retrouve les problèmes pointés par Foucault, les *a priori* formels prenant le pas sur les *a priori* historiques.

---

[1] Calenge 1994, p. 401 : « Gérer, développer et renouveler une collection est l'objet même du métier de bibliothécaire : toutes les sciences appelées à la rescousse (sociologie, linguistique, psychologie...), toutes les techniques extérieures (architecture, informatique...) ou développées de façon interne (catalogage, indexation...), tournent autour de cet objet unique qui légitime le bibliothécaire et constitue le cœur de son activité : la collection vivante et en action ».

A - Outils utilisés pour la sélection des sources

Calenge [1994] propose[1] deux types d'outil pour guider les acquéreurs dans leur tâche de sélection des sources : d'une part, la classification décimale Dewey, d'autre part, une échelle (à établir par chaque établissement) codant les usages possibles d'un document (en fonction de leur « complexité »)[2].

Sans revenir ici sur les critiques dont a pu faire l'objet la classification Dewey en matière de représentation des domaines de connaissance, on peut s'étonner de voir intervenir, dès le choix des sources, un langage documentaire.

Trois types d'arguments sont avancés par Calenge. Le premier tend à justifier la classification Dewey au regard de l'utilisation d'un autre langage documentaire[3]. Les deux autres mettent en avant des aspects pragmatiques : la classification Dewey constitue le « langage commun » des professionnels[4] ; elle est par ailleurs utilisée dans les catalogues des éditeurs comme par les intermédiaires qui diffusent les nouveautés éditoriales[5].

Le choix d'un langage documentaire pour trier parmi les sources existantes n'est donc pas à proprement parler problématisé :

- on ne tient pas compte du fait que la classification Dewey, si elle « synthétise le contenu » d'un document, n'est pas un outil neutre. Autrement dit, les stratégies de choix propres aux professionnels qui semblaient être mises en avant par la notion de « politique d'acquisition », sont ici réduites au silence : on s'en remet à la notoriété d'un instrument comme si les bibliothécaires n'en avaient pas été les agents actifs ;

- on ne distingue pas les deux stratégies d'exploration des sources et d'exposition des documents : en recourant au même outil pour trier les sources et pour classer les documents, ce qui devait constituer en propre le métier du bibliothécaire, la mise en collection des documents, ne trouve plus ici de moyens spécifiques.

On le voit : les règles d'exploration des sources, telles qu'elles sont habituellement établies en bibliothèques, ne sortent pas d'une approche lexicale du contenu. La même perception des mots et des choses qui caractérise, dans la démarche classique, l'indexation d'un document[6], se retrouve, sous une forme identique, lorsqu'il s'agit de formaliser les critères de choix d'une source. La distinction source/document n'est pas établie ni les stratégies dévoilées.

---

[1] Calenge s'appuie pour ce faire sur l'analyse de chartes documentaires existantes.

[2] Ce second outil (évaluant les « niveaux de lecture » d'un document : adulte, enfant, etc.) n'est pas utilisé pour sélectionner à proprement parler les sources documentaires ; il opère plutôt un filtrage sur un ensemble de sources sélectionnées à partir de la classification Dewey. C'est pourquoi nous n'aborderons pas ici ce deuxième type d'outil.

[3] Voir Calenge 1994, p. 129 : « C'est le principe décimal des classifications qui reste un atout dans la définition et l'analyse des objectifs documentaires car il autorise des regroupements exhaustifs interdits aux descripteurs alphabétiques (liste Rameau ou thésaurus spécialisés) ».

[4] *Ibid.*, p. 124 ou p. 134, entre autres.

[5] *Ibid.*, p. 396.

[6] Voir notre première partie : les problèmes théoriques de l'indexation.

Ne sont pas non plus prises en compte les particularités des langages documentaires. Comme l'ont établi Bourion et Malrieu [1994], sur la base d'une étude d'un plan de classification en psychologie sociale, on ne saurait ignorer les « instructions de lecture » implicitement à l'œuvre dans un langage documentaire. Nous renvoyons, pour le détail, à l'analyse[1] qu'elles mènent pour exhiber ce qu'elles nomment le « discours classificatoire[2] » des indexeurs, en adoptant une approche à la fois cognitive et linguistique. Les résultats de leur analyse rejoignent ceux de Sylvie Bruxelles, précédemment évoqués[3], sur les relations d'intertextualité qui s'établissent entre plans de classement et textes à indexer. Ce que ces auteurs montrent des contraintes de lectures exercées par les langages documentaires sur le choix des mots nous semble être tout aussi actif dans le choix des sources.

En outre, il importe de rappeler ce qui est régulièrement observé par les professionnels eux-mêmes. Le rangement d'une source sous un indice de classification est toujours le résultat d'une distorsion, plus ou moins forte, et au final toujours le résultat d'un choix[4] : les sources se classent difficilement elles-mêmes dans le cadre d'une représentation des connaissances établie sur des bases depuis longtemps perdues et oubliées[5], modifiée de façon *ad hoc* au gré des difficultés rencontrées. De là, un certain nombre de « licences » que s'accordent nécessairement les documentalistes et les bibliothécaires, sans toujours justifier ou simplement expliciter les libertés prises par rapport au référentiel de base.

Un exemple parmi les nombreux cas que l'on trouve dans les pratiques documentaires est relevé par Bourion et Malrieu[6] : le CNRS dispose de deux bases de données qui distinguent, d'une part, « les sciences humaines, sociales et économiques » (base Francis) et, d'autre part, « les sciences, les technologies et la médecine » (base Pascal). Il faut connaître les péripéties de l'histoire de l'institution pour savoir que le domaine de la psychologie ne fait pas partie, dans les bases de données du CNRS, du domaine des sciences humaines mais de celui des sciences de la vie (base Pascal).

Là encore, ce qui vaut pour les pratiques de classement vaut aussi pour les pratiques d'exploration des sources, puisque ce sont les mêmes outils qui sont employés. Dans le cadre d'une organisation du travail en général très parcellisée comme il est courant dans les bibliothèques d'importance, un acquéreur n'a de vue sur son domaine que ce que lui propose un indice Dewey : sa stratégie d'exploration se trouve de fait très limitée. C'est ainsi que si d'aventure au sein d'une base de données, deux documents relevant d'indices Dewey différents venaient à être rapprochés, ce serait là un pur hasard, un fait de rencontre qui ne devrait rien à une stratégie de regroupement. C'est en ce sens qu'il paraît très délicat de pouvoir *a posteriori*, comme le suggérait Suzanne Bertrand-Gastaldi, rétablir, par le biais de

---

[1] Bourion et Malrieu 1994, p. 83-131 : elles étudient le plan de classement en tant que « genre textuel » de façon à cerner les contraintes qu'il fait peser dans la construction de l'interprétation.
[2] Nous y revenons ci-après, § III.2.1.
[3] Bruxelles 1991, voir chapitre II § III.1.
[4] Bourion et Malrieu 1994, p. 86 : « Mais les "objets" à classer (les articles) renvoient à des concepts qui peuvent nécessiter la référence à différentes spécialités, se situer à la périphérie des courants dominants parce qu'ils s'autodéfinissent au travers de concepts et "territoires nouveaux", en émergence ».
[5] Ainsi de la classification décimale Dewey, établie en 1873 sur la base de la littérature alors disponible dans les bibliothèques américaines d'Amherst, des régions de New-York et New England. Un historique peut être trouvé dans Comaroni 1988 par exemple.
[6] Bourion et Malrieu 1994, p. 86.

« formations discursives », des connexions entre documents. Selon nous, il importe de desserrer, dès le niveau de la sélection documentaire, l'étau des langages documentaires.

B - Outils utilisés pour la sélection des objets à indexer

Dans le cadre de notre expérimentation (où l'on cherchait à définir ce qui guidait la sélection des objets d'indexation au sein d'une source considérée en soi comme pertinente), si l'on retrouve le même recours à des référentiels établis sur des critères de contenu, on trouve aussi d'autres types de règles.

De façon générale, les organismes documentaires étudiés s'appuient, pour établir leur sélection, soit sur une liste de thèmes (par exemple : la communication, les arts, les médias, la vie culturelle, etc.), soit sur les rubriques du journal *Le Monde* : pour un organisme documentaire spécialisée dans l'économie par exemple, seules les pages ainsi intitulées seront explorées.

Ce recours à un mode de guidage émanant de la source elle-même peut être, dans certains cas, problématique. Sur ce point, un entretien mené sur la sélection documentaire exercée par un documentaliste utilisant les règles d'exploration par rubriques nous a montré le caractère réducteur de ce type d'exploration. Habitué à ne consulter que les pages « Communication » du journal, le documentaliste n'avait pas retenu un article qui, classé dans la rubrique « Économie », aborde le domaine des technologies nouvelles habituellement traité par le service de documentation : l'article en question[1] (« La Route intelligente ») rend compte du premier congrès sur la télématique des transports. Le documentaliste, interrogé sur la non-sélection de cet article, reconnaissait la pertinence de cet article pour les domaines qu'il couvre, tout en hésitant à le retenir à cause de son apparition dans la rubrique « Économie ».

Sur ce point, le système des rubriques dans le journal *Le Monde* est perçu de la même façon que le système des indices de la classification Dewey : les deux sont appréhendés comme des référentiels objectifs et nécessairement valides, sans que la particularité de leur construction ne soit prise en compte. Si nous ne remettons pas nécessairement en cause le recours à des référentiels établis, nous remettons en cause la perception dont ils font l'objet. Il semble en effet nécessaire de les considérer comme porteurs d'« instructions de lecture », pour reprendre les termes de Bourion et Malrieu, qui peuvent intéresser le lecteur à ce titre-là : c'est alors le point de vue du *Monde* sur les événements qui peut être retenu et indexé comme tel et non les événements en eux-mêmes.

Sur ce point se distinguent, parmi les organismes documentaires consultés, deux types de perception différente qui déterminent deux types de stratégie :

(i) pour ceux qui considèrent *Le Monde* comme une source d'information comme une autre, le quotidien sera lu au travers d'un référentiel basé sur le contenu (liste thématique externe ou système de rubriques interne au journal) ;

(ii) pour ceux qui considèrent *Le Monde* comme source spécifique dans la presse française, les stratégies d'exploration seront moins déterminées par un contenu supposé : les signatures des articles pourront, par exemple, prendre

[1] *Le Monde*, 1/12/1994, p. 19.

plus d'importance que le contenu traité. Perçu sous l'angle du « journal de référence » (l'expression revient souvent dans les entretiens), le journal, dont il semble important d'avoir le point de vue, sera alors exploré de façon large et selon des règles qui restent alors très implicites.

Il nous semble que l'on peut voir, dans ce dernier type d'exploration qui cherche à prendre en compte la spécificité de la source, les prémices d'une démarche propre à intégrer la notion de formation discursive établie par Foucault.

De la même façon que, précédemment dans le point A, apparaissaient, au-delà de l'adoption de règles explicites, certaines zones de liberté prises par les acquéreurs (par rapport aux plans de classement), se dessinent également, dans le cadre de notre expérimentation, quelques décalages entre principes de sélection avoués et principes de sélection appliqués.

Ainsi, le secteur « presse » de *La Documentation française* met-il en avant, dans ses déclarations de principe, « la pluralité des sources [qui] apporte à l'utilisateur une information équilibrée, complète et impartiale[1] ». En effet, le secteur « presse » de *La Documentation française* est l'un des rares services d'information publics qui traitent la presse de tous les partis politiques (De *Présent* à *Révolution* en passant par *Rouge et Vert*) comme de tous les syndicats (*FO hebdo*, *CFDT magazine*, etc.). À partir de ces sources, les critères de sélection explicites sont les suivants[2] :

- « exhaustivité sur l'organisation et l'activité des pouvoirs publics ;
- large sélection sur la vie des partis, des syndicats, du patronat ;
- sélection rigoureuse quant au contenu informatif sur les questions d'actualité liées au débat politique (problèmes de société, grandes entreprises, vie des médias, nouvelles technologies, etc.) ».

On devrait logiquement s'attendre à ce que la sélection exercée sur *Le Monde* soit finalement assez « rigoureuse » puisque sont abordées essentiellement dans ce journal des « questions d'actualité liées au débat politique ». Or, pour l'année 1994, les références proposées aux utilisateurs sont extraites pour un tiers (26%) du *Monde*, qui arrive en tête du palmarès des sources présentées dans la base de données de *La Documentation française*. À eux trois, les quotidiens *Le Monde*, *Le Figaro* et *Libération* représentent, en 1994, 60% des articles proposés pour rendre compte de la vie politique française. Il n'y a là, sans doute, rien de très étonnant quand on connaît le poids de chacun de ces quotidiens dans le paysage de la presse française ; reste que les règles explicites de sélection ne sont pas celles qui sont appliquées.

Une analyse de type sociologique pourrait sans doute contribuer à faire émerger les règles implicites effectivement appliquées.

Sous l'angle de vue que nous avons retenu, on essaiera de dégager les principes de sélection qui peuvent être mis en œuvre, en reprenant les études que d'autres chercheurs ont pu mener sur un plan plus linguistique.

---

[1] *Infos-Bipa*, mise à jour 1995, [p. 1].
[2] *Ibid.*, [p. 2].

II.2.2.2 - Les principes implicites d'exploration des sources

À la suite de Suzanne Bertrand-Gastaldi [1989], qui reprend elle-même les conclusions de chercheurs anglophones comme Beghtol [1986], on peut dégager un principe général d'exploration des sources en indexation : le principe de l'intertextualité, actif à plusieurs niveaux que nous indiquerons succinctement.

Les théories de l'intertextualité sont nombreuses, diverses, souvent prises en défaut pour celles qui insistent sur le rapprochement entre textes effectué sur la base du seul contenu[1]. Sans pouvoir entrer dans les problématiques spécifiques à ces théories, nous proposons de considérer l'intertextualité sous un angle qui insiste sur la transformation des usages d'un texte, lorsqu'il passe d'un ensemble de textes à un autre. C'est pourquoi nous retiendrons l'approche suivante de l'intertextualité : « [L'intertextualité] absorbe l'énoncé qu'elle emprunte à un modèle antérieur pour l'inscrire dans un autre ensemble textuel : elle ne se contente pas toutefois de l'incorporer, elle le soumet à une activité transformatrice, elle enchâsse le texte primitif dans un contexte nouveau dans le dessein d'en modifier le sens. L'intertextualité ne recouvre ainsi pas uniquement une opération mémoriale et assimilatrice, elle n'est pas uniquement une transplantation d'un texte dans un autre, mais elle se définit par un travail d'appropriation et de réécriture qui s'applique à recréer le sens, en invitant à une lecture nouvelle.[2] »

Le principe de l'intertextualité ainsi compris met en valeur ce que nous avons appelé l'espace discursif propre à l'indexation, l'espace d'interprétation des sources, réalisé bien avant l'assignation ou l'extraction effectives de descripteurs.

L'intertextualité, la mise en rapport de textes dans le but de « construire un univers relationnel[3] », se réalise sur plusieurs plans dans la phase de sélection des sources :
A - entre les sources elles-mêmes ;
B - entre les documents déjà sélectionnés et les sources explorées ;
C - entre les usages antérieurs des sources et les sources explorées.

A - Type 1 d'intertextualité : au niveau des sources

Cet aspect de l'intertextualité est surtout visible dans les cas de sélection d'objets d'indexation parmi un ensemble de sources finies (c'est le cas de notre expérimentation qui porte sur le dépouillement d'une source retenue *a priori*). La décision qui préside au choix de tel ou tel article du *Monde* comme au choix de son insertion dans une série (revue de presse, par exemple) se prend au regard des autres sources (d'autres quotidiens par exemple), c'est-à-dire sur la base d'un jugement qui repose sur le « déjà-dit », sur la base d'un contenu jugé semblable. De là deux types de stratégie : soit le « déjà-dit » apparaît comme une redondance et la source consultée n'est pas retenue ; soit il apparaît au contraire comme présentant un point de vue complémentaire et la source est alors sélectionnée à ce titre. Les deux cas de figure se trouvent dans notre enquête. À partir d'un même jugement de ressemblance (notion de « déjà-dit »), deux types de stratégies différentes sont adoptées sur la base de règles qui restent implicites.

---

[1] Marandin [1979] et [1993] ainsi que Pêcheux [1975] ont pu mener la critique sous cet angle.
[2] Eigeldinger 1987, p. 11, cité *in* Bertrand-Gastaldi 1993, p. 145.
[3] Ricardou cité *in* Bertrand-Gastaldi 1989, p. 142.

B - Type 2 d'intertextualité : au niveau des documents

Comme nous l'avons précédemment évoqué, les documents déjà constitués pèsent sur le choix des sources, soit de façon directe soit de façon indirecte quand la sélection des sources se mène par le biais d'un langage documentaire : comme le note Bertrand-Gastaldi, « l'interprétation des nouveaux textes subit donc en partie l'influence des textes antérieurs[1] ». Là encore, si c'est un même type d'intertextualité qui est mis en œuvre, les décisions prises peuvent être opposées. Dans certains cas, seules les sources présentant une « nouveauté » par rapport à l'existant seront constituées comme documents ; dans d'autres cas, seules les sources faisant « suite » aux documents en place seront retenues. Là encore les règles qui président à ces choix ne sont pas explicitées. À l'analyse, apparaissent des stratégies qui, pour être établies sur le même constat, aboutissent pourtant à des décisions opposées.

C - Type 3 d'intertextualité : au niveau des usages antérieurs

L'utilisation faite des sources dans des contextes qui peuvent être extra-documentaires n'est pas non plus sans exercer une influence sur la sélection ou le rejet d'une source : là encore, les décisions prises à partir de constats semblables peuvent être opposées et les stratégies restent non dévoilées.

Odile Le Guern, travaillant sur le traitement documentaire de l'image, relève que le droit d'accès d'une image dans une iconothèque peut être contraint par des usages antérieurs : « Il est parfois nécessaire de tenir compte du contexte que constitue le réseau des utilisations précédentes, des précédents discours qui ont intégré les documents. [...] Certaines images sont devenues des images symboles du premier discours qui les a intégrées et ne pourront que très difficilement faire l'objet d'autres lectures pour d'autres utilisateurs, originales par rapport à ce premier discours.[2] »

Si certains usages bloquent la transformation d'une source en document, inversement, les usages extra-documentaires d'une source peuvent décider de son intégration *a posteriori* dans l'espace documentaire. C'est le cas en particulier pour les indexeurs travaillant sur la presse d'actualité, l'entrefilet d'un jour pouvant, quelques semaines plus tard, se lire comme le début de toute une chaîne d'événements, méritant alors d'être intégré dans la collection documentaire.

L'existence documentaire d'une source n'est donc pas intrinsèquement liée à un type de texte. C'est ainsi que tous les organismes documentaires interrogés dans notre enquête, sans exception aucune, gardent une collection complète du journal *Le Monde*, au moins sur une année, malgré le travail de dépouillement, parfois très long (pour certains, près de deux heures consacrées uniquement à la sélection) dont il a fait l'objet. Certains d'entre eux[3] souscrivent parallèlement un abonnement à la base de données du *Monde*, pour éventuellement récupérer des articles anciens qui n'auraient été pas retenus. En somme, on retrouve, malgré tous les discours sur la valeur d'échange des réseaux, sur l'impossibilité de l'exhaustivité ou, plus positivement, sur la nécessité de se doter d'une politique d'acquisition, la stratégie du « au cas où », dominante dans la profession.

---

[1] Bertrand-Gastaldi 1989, p. 147.
[2] Odile Le Guern 1989, p. 428.
[3] Cinq organismes documentaires sur les neuf interrogés sur cette question (le service de documentation du *Monde* n'est pas ici comptabilisé).

Nous y voyons l'absence de stratégie explicite d'exploration des sources qui ne présente pas le seul inconvénient de générer une forte déperdition d'« énergie » professionnelle, mais, plus fondamentalement, qui enlève au choix des sources en indexation toute possibilité de systématicité et de maîtrise.

II.2.3 - CONCLUSIONS INTERMÉDIAIRES

En posant comme horizon théorique la notion de « système-archive » proposée par Foucault, nous voulions mettre en valeur :

- d'une part, que les pratiques d'indexation mettent en œuvre des moyens pour réaliser une « transformation discursive », pour établir un niveau proprement documentaire, où un texte peut se concevoir à la fois sous l'angle d'une source (événement) et sous l'angle d'un document (chose). Nous avons sur ce point, par quelques exemples, montré la diversité des mises en discours qui s'observe sous l'adoption d'un même terme d'indexation, ainsi que les traces qu'un document pouvait garder de la source d'où il provient ;

- d'autre part, que cette « transformation discursive » reste, dans la majeure partie des cas, déréglée : soit parce que les règles explicites qui guident l'exploration des sources en indexation s'appuient sur des langages documentaires et reproduisent ce faisant une « analyse de contenu » dont on a montré précédemment les limites ; soit parce que les règles implicites qui sont mises en œuvre à partir du principe de la confrontation textuelle ne reposent sur aucune stratégie réellement maîtrisée.

Cependant, certains types d'exploration des sources en indexation, notamment ceux qui prennent en compte la spécificité du foyer énonciatif, laissent penser qu'il n'est pas complètement absurde d'envisager les règles d'exploration des sources en indexation dans les termes des formations discursives de Foucault. Une étude précise mériterait d'être menée sur ce point.

Au demeurant, il apparaît que la stratégie d'exploration des sources en indexation telle qu'elle se mène habituellement, si elle reste le plus souvent implicite, n'en produit pas moins des espaces documentaires nécessairement différents d'une pratique documentaire à l'autre. Ce n'est pas cette différence qui nous paraît problématique ; nous estimons au contraire que la valeur d'un système d'information tient au regard qu'il porte ou qu'il permet de porter sur une production éditoriale. Ce qui nous paraît problématique, c'est que la spécificité de l'espace documentaire créé ne soit pas systématiquement prise en compte, ni au moment de la constitution de la collection, ni au moment de l'attribution des descripteurs. Or, ce n'est qu'en ayant conscience que la sélection des sources aboutit à la création d'un univers documentaire particulier, que l'on peut véritablement poser la problématique de l'usage des mots en indexation : les mots en indexation ont alors pour fonction d'établir un passage entre l'espace documentaire des indexeurs et l'espace d'usages des utilisateurs. En ce sens, tout comme une stratégie d'exploration explicite des sources permet d'avoir une maîtrise de l'espace documentaire que l'on crée, une stratégie d'exposition des documents s'avère nécessaire pour maîtriser le passage d'un espace à l'autre. C'est sur cette problématique que nous nous interrogeons dans le paragraphe suivant.

III - Stratégie d'exposition des documents en indexation

La notion de discours documentaire, entendue comme espace d'organisation spécifique des documents, si elle interroge l'indexation sous l'angle du choix des sources en la contraignant à expliciter ses stratégies d'exploration, permet également de dégager une autre problématique, maintenue elle aussi le plus souvent sous la seule autorité du langage documentaire : comment rendre l'espace documentaire accessible aux utilisateurs ?

Avec la notion de discours documentaire telle que nous essayons de la construire, la problématique de la communication en indexation ne se pose plus, ou plus uniquement, sous l'angle du mot juste ; elle appelle désormais la dimension plus large du discours et de ses conditions d'interprétation. Si, au travers de règles plus ou moins explicites, problématisées et systématiques, l'indexation parvient à transformer des sources hétérogènes en ensembles de documents révélant une certaine cohérence (ne serait-ce qu'« optique » pour reprendre la formulation de Latour), le choix de telles ou telles unités linguistiques pour signaler des liens entre documents peut, au sein du cadre restreint des indexeurs, bénéficier d'une certaine stabilité : mais que devient cette stabilité quand le cercle des locuteurs s'élargit jusqu'aux utilisateurs ? Se dessine la nécessité, pour l'indexation, de disposer d'une stratégie d'exposition des documents qu'elle a créés.

Pour aborder l'aspect de l'indexation sous l'angle d'une stratégie d'exposition des documents, nous procéderons de la façon suivante :

- d'une part, nous disposerons un cadre théorique qui nous permette de problématiser la question des différents « espaces » de locuteurs en jeu dans l'indexation, « espace » des indexeurs et « espace » des utilisateurs ; à ce titre, on empruntera à Kripke son approche des mondes possibles. Cet emprunt, quoique nécessaire, est problématique : issu de la logique, le modèle des mondes possibles, s'il peut être utilisé dans une perspective linguistique, nécessite des reformulations, qui flirtent toujours, dangereusement, avec l'emprunt métaphorique. Nous essayerons d'expliciter ce point et les conditions dans lesquelles nous pourrons utiliser ce concept dans le cadre de notre recherche ;

- d'autre part, compte tenu de la problématique posée, nous présenterons quelques moyens qu'utilise ou que pourrait utiliser l'indexation pour établir des « ponts » entre les différents mondes possibles qu'elle doit traverser.

III.1 - La notion de « monde possible » dans le modèle de Kripke

Nous avons tenté de montrer précédemment que la pratique d'indexation exerçait une « mise en discours » des sources au terme de laquelle était créé un espace documentaire. Nous avons utilisé plusieurs formulations pour rendre compte de cette création des documents en indexation[1]. Pour pouvoir poursuivre notre recherche, il nous paraît désormais nécessaire d'essayer de préciser la terminologie employée.

[1] Espace de représentation, espace de discours, univers d'interprétation par exemple.

Le recours aux images d'« univers », d'« espace » ou de « monde », pour faire référence à un domaine d'interprétation, est courant[1]. Emprunter le concept logique établi par Kripke et examiner les conditions de son emploi dans une perpective linguistique nous paraissent être de nature à rendre opératoire l'utilisation de ces images, du moins dans le cadre de notre recherche. Le concept de « monde possible » nous permettra notamment de traiter la question du conflit des interprétations entre groupes de locuteurs (indexeurs et utilisateurs) sans poser l'enjeu dans les termes d'objectivité ou de subjectivité, comme il est souvent fait dans la littérature classique sur l'indexation[2].

III.1.1 - PRÉSENTATION DE LA NOTION DE « MONDE POSSIBLE »

La notion de monde possible proposée par Kripke constitue une « version laïque » de la notion originellement formulée par Leibniz. Si Leibniz oppose le monde réel, créé par Dieu, aux mondes possibles, laissés intentionnellement inactualisés par Dieu, pour Kripke, outre que le monde réel est une sorte de monde possible, le monde possible est toujours une création humaine : « Un monde possible n'est pas un pays lointain qu'on rencontre sur son chemin ou qu'on regarde au télescope. [...] Un monde possible est *donné par les conditions descriptives que nous lui associons.* [...] Les "mondes possibles" sont *stipulés*, ils ne sont pas *découverts* par de puissants télescopes.[3] »

Cette thèse de Kripke s'oppose aussi à celle d'autres logiciens, qui pensent la notion de « mondes possibles » dans les termes d'une « théorie des répliques[4] ». Cette théorie, qui suppose qu'il existe un univers de référence dont on peut imaginer des variantes, se fonde sur la notion d'« identité » entre mondes possibles ; or c'est précisément ce point que met en cause Kripke. En effet, comme nous l'avons précédemment évoqué, sa description de la référence se veut entièrement déconnectée d'une description qualitative, en termes de traits définitoires nécessaires et suffisants : « Même s'il y avait un ensemble purement qualitatif de conditions nécessaires et suffisantes pour être Nixon, la conception que je défends ici n'impliquerait pas qu'il faille trouver ces conditions *avant* d'être en mesure de demander si Nixon aurait pu gagner les élections, ni qu'il faille reformuler la question en termes de telles conditions. Nous pouvons considérer simplement *Nixon* et demander ce qu'il aurait pu lui arriver à *lui* si diverses circonstances avaient été différentes.[5] »

Cet exemple montre que, dans tous les mondes possibles, Nixon reste un être humain : autrement dit, la création de mondes possibles est contrainte. Tout contexte imaginable ne peut fonctionner dans tous les cas comme un monde possible ; il doit, pour constituer un « monde possible », se révéler contrefactuel[6]. On doit pouvoir opposer ce qui est possible à ce qui ne l'est pas, c'est-à-dire cerner l'« idée de possibilité qui est en jeu[7] » dans les mondes possibles créés. Ainsi Kripke propose-t-il de considérer un monde possible comme une « situation

---

[1] On la retrouve par exemple chez Ricœur [1971, p. 175-187] ou chez Rastier [1994].
[2] Par exemple Quinn 1994 pour une synthèse.
[3] Kripke 1982 [1972], p. 32.
[4] *Ibid.*, p. 33-34.
[5] *Ibid.*, p. 35.
[6] « Si la proposition *Aristote aimait les chats* est tenue pour factuelle, la proposition *Aristote n'aimait pas les chats* est tenue pour contrefactuelle ». Notons que c'est cette possibilité de construire P et non-P à partir du nom propre Aristote qui prouve que le nom propre « n'est pas le sténogramme d'un paquet de prédicats identificateurs ». Milner 1989, p. 331.
[7] Kripke 1982 [1972], p. 34.

contrefactuelle », un « état » et non une entité concrète : « Il suffit de décrire en quoi la "situation contrefactuelle" diffère (de façon pertinente) des faits réels ; on peut concevoir la "situation contrefactuelle" comme un mini-monde ou mini-état, restreint aux aspects du monde qui sont pertinents pour ce qui est en question.[1] »

Selon cette approche des mondes possibles intentionnellement créés, l'« identité » observable entre mondes possibles est une conséquence et non une cause : « Les théoriciens ont souvent dit que nous identifions les objets à travers les mondes possibles grâce à leur ressemblance avec l'objet de départ sous les aspects les plus importants. [...] Au contraire, nous commençons avec les objets, que nous avons et que nous pouvons identifier dans le monde réel. Nous pouvons ensuite nous demander si certaines choses auraient pu être vraies de ces objets.[2] »

Rejetant l'existence *a priori* des mondes possibles et, de ce fait, la nécessité de les décrire « qualitativement », Kripke propose une approche de la notion de mondes possibles qui met en avant deux aspects : la création contrainte des mondes possibles et le caractère résultatif de l'« identité » des objets.

Ces deux aspects sont intimement liés : les « mondes possibles » n'existent pas en soi mais sont toujours créés à partir de la perception d'un objet dont on se demande s'il pourrait être doté de propriétés que l'on ne lui connaît pas dans le monde où l'on se trouve. La démarche de création d'un monde possible ne suppose donc pas une recherche d'identité de propriétés d'un objet ; au contraire, elle stipule que les propriétés d'un objet peuvent changer. Autrement dit, pour Kripke, les critères d'identification d'un objet sont toujours relatifs à un monde : ce ne sont pas les mêmes critères que l'on met en œuvre pour distinguer un objet dans le « monde réel » et dans un « monde possible[3] ». C'est ainsi que, pour Kripke, il ne relève pas d'une théorie des mondes possibles de se demander si tel objet du monde réel (la reine d'Angleterre, dans son exemple[4]) aurait pu être un autre objet (un cygne, dans son exemple) dans un monde possible : ce type de question repose sur l'identification inter-mondes de propriétés semblables d'objets différents et sur l'existence *a priori* des mondes possibles. C'est en ce sens que la construction des mondes possibles est, dans le modèle de Kripke, contrainte. Les contraintes qui pèsent sur la création des mondes possibles sont exprimées en termes d'« essences[5] » et relèvent de trois ordres[6] :

- l'origine, par exemple « être humain » pour la reine d'Angleterre : autant on peut créer un monde où la reine d'Angleterre serait une pauvresse, autant on ne peut pas créer un monde où elle serait un cygne ;

- la matière, par exemple « être en bois » pour une table en bois : on peut créer un monde où la table en bois dont on parle dans le monde réel est dans un

---

[1] Kripke 1982 [1972], p. 170.
[2] *Ibid.*, p. 41.
[3] *Ibid.*, p. 37 : « Les propriétés qu'un objet a dans tout monde contrefactuel n'ont rien à voir avec les propriétés dont on se sert pour l'identifier dans le monde réel ».
[4] *Ibid.*, p. 100-102.
[5] *Ibid.*, p. 41 : « Certaines propriétés d'un objet peuvent lui être essentielles, dans la mesure où il n'aurait pas pu ne pas les avoir. Mais ces propriétés ne servent pas à identifier l'objet dans un autre monde possible, car une telle identification n'est pas requise ».
[6] Kripke n'établit pas de cette façon les propriétés essentielles des objets : ses propos sont plus diffus, complexes, nuancés ; pour simplement donner quelques illustrations de ce que semble vouloir dire Kripke sur les propriétés essentielles, on reprend, sans les discuter, la typologie et les commentaires que propose Engel 1985, chap. V.

autre endroit que celui que l'on perçoit ; on ne peut, en revanche, imaginer qu'elle puisse fondre (c'est-à-dire imaginer qu'elle soit en glace par exemple) ;

- la forme, par exemple « avoir la forme de table » pour une table : comme précédemment, on ne peut créer un monde possible, au sens où Kripke l'entend, dans lequel une table serait un vase.

Pour autant, ces propriétés essentielles, précise Kripke, ne répondent pas à la question : « quelles propriétés un objet doit-il garder pour ne pas cesser d'exister ?[1] », mais plutôt à celle-ci : « quelles propriétés (atemporelles) un objet n'aurait-il pas pu ne pas avoir ?[2] ». La différence entre les deux types de question renvoie à l'existence *a priori* ou pas des mondes possibles : si les mondes possibles sont perçus comme existants *a priori*, ils révèlent une temporalité propre ; pour Kripke, les mondes possibles, n'existant pas *a priori* mais étant construits à partir du « monde réel » (le monde d'où l'on parle, pourrait-on dire), s'introduisent *dans la temporalité*, dans l'« histoire » du monde réel : « D'habitude, lorsque nous nous demandons si, intuitivement, quelque chose aurait pu arriver à un objet donné, nous nous demandons si l'histoire de l'univers aurait pu se dérouler comme elle s'est effectivement déroulée jusqu'à un certain instant, et adopter ensuite un cours différent du cours réel, de sorte que, à partir de là, les vicissitudes de cet objet auraient pu être différentes de celles qu'elles ont été.[3] »

C'est pourquoi, selon Engel, les propriétés essentielles de Kripke ne constituent pas des propriétés individualisantes, des « essences individuelles ». Elles relèvent plutôt d'essences « sortales » qui « n'ont pas pour fonction d'*individualiser* un individu en lui attribuant une propriété, mais seulement de fixer sa référence en le rapportant à une espèce. On s'assure ainsi que César n'est pas un chien mais un homme[4] ».

La théorie des mondes possibles telle que Kripke la conçoit nous paraît particulièrement pertinente pour notre étude, dans le sens où, d'une part, elle insiste sur le caractère construit et contraint des mondes possibles et où, d'autre part, elle repositionne la question de l'identification des objets entre mondes. Ces deux points sont essentiels pour l'approche de l'indexation que nous proposons.

Avant d'en venir à l'utilisation que nous ferons de la théorie kripkéenne des mondes possibles, il nous faut déterminer dans quelle mesure elle peut être appliquée à des objets linguistiques (des textes, des énoncés, des discours).

III.1.2 - INTERPRÉTATION LINGUISTIQUE DE LA NOTION DE « MONDE POSSIBLE »

Sans souci d'exhaustivité, nous avons relevé dans la littérature que nous avons explorée pour cette recherche, plusieurs types « d'utilisations linguistiques » de la notion de mondes possibles. Ainsi, sont assimilés à des mondes possibles des textes

---

[1] Kripke 1982 [1972], p. 103, note 57.
[2] *Id.*
[3] Kripke 1982 [1972], p. 104, note 57.
[4] Engel 1985.

(la plupart du temps, ce sont les romans[1] qui sont appréhendés en termes de mondes possibles) ou encore des discours[2].

Gary-Prieur [1994] ainsi qu'Eco [1985 (1979)] s'attachent particulièrement à justifier de telles utilisations du concept de Kripke : leurs argumentations, si elles conduisent à de semblables conclusions, ne portent pas sur les mêmes éléments.

Gary-Prieur [1994] met en avant que les mondes possibles de Kripke sont « stipulés » et qu'on peut entendre par là qu'ils « n'ont pas d'existence extérieure au discours qui les met en place[3] » ; autrement dit, c'est par le discours que se créent les mondes possibles. Gary-Prieur envisage la notion de monde possible sous l'angle du créateur de ce monde, du côté du locuteur. Elle reste ce faisant très proche des conceptions de Kripke : en effet, comme le souligne l'un de ses commentateurs[4], les mondes possibles sont, pour Kripke, des « manières de parler », de pures entités linguistiques[5].

Insistant sur un autre aspect de la théorie de Kripke – la notion de situation contrefactuelle –, Eco propose une autre approche des mondes possibles qui ne se situe plus du côté du locuteur mais du côté de l'interlocuteur, pourrait-on dire, du lecteur. Eco considère un monde possible comme un fait d'interprétation : si sa position rejoint celle des auteurs qui abordent le roman en termes de monde possible, elle prend quelque distance avec les propos de Kripke.

Selon Eco, la lecture d'un texte progresse par formulations successives d'hypothèses, de prévisions établies par rapport à ce que le texte a préalablement disposé : un lecteur stipule « un cours d'événements possible ou un état de choses possibles » qui se verront confirmés ou infirmés par la poursuite de la lecture[6]. Dans ce cadre, un monde possible est une prévision que le lecteur construit sur la base de ce que dit le texte, qui constitue alors une contrainte sur la création des mondes possibles. Eco insiste sur le fait que l'interprétation d'un texte ne relève pas d'un choix entre alternatives mais d'un choix entre possibilités, définies par le texte lui-même[7]. En ce sens, il propose de définir la notion de monde possible comme « un état de choses exprimé par un ensemble de propositions où, pour chaque proposition, soit *p* soit non-*p*[8] ». Sur ce point, Eco rejoint en partie Gary-Prieur, bien que leur cadre d'approche soit radicalement différent (sémiotique des textes narratifs pour Eco, grammaire du nom propre pour Gary-Prieur) : c'est à l'intérieur d'un énoncé que peuvent intervenir plusieurs mondes possibles[9].

Compte tenu des caractéristiques que Kripke attribue à la notion de monde possible (construction contrainte), il semble légitime de pouvoir associer la notion de monde

---

[1] Cf. Corblin 1995, p. 198 et suiv. pour des exemples : « Un roman stipule un monde possible », ou encore Eco 1985 [1979], chapitre 8, p. 157-225 : « Structures de mondes », ou Gary-Prieur 1994, p. 26, note 1.
[2] Fradin et Marandin [1979] parlent de « reformulation » du concept de Kripke pour étudier les aspects de la rigidité comme effets de référence dans le discours.
[3] Gary-Prieur 1994, p. 21.
[4] Engel 1985, chapitre IV.
[5] *Id.* : « Kripke semble se ranger dans le camp de ceux qui considèrent le langage en termes de mondes comme heuristique et réduisent ceux-ci à de pures entités linguistiques ou des manières de parler ».
[6] Eco 1985 [1979], p. 146.
[7] *Ibid.*, p. 160.
[8] *Ibid.*, p. 165.
[9] Gary-Prieur 1994, p. 22.

possible à la notion linguistique de construction référentielle : un monde possible correspond alors à la construction de la référence en discours. C'est sur cette base que nous nous appuierons pour élaborer notre propre approche de la notion de monde possible.

III.1.3 - ENJEU DE LA NOTION DE « MONDE POSSIBLE » EN INDEXATION

Cet essai d'interprétation des enjeux de l'indexation dans le cadre de la théorie de Kripke a pour objectif de repositionner la problématique des mots en indexation au regard de la question de l'identité de l'objet telle que la pose Kripke. Il s'agit de montrer que les descripteurs ne peuvent avoir pour rôle d'identifier le « même » objet dans tous les mondes possibles, mais qu'ils ont plutôt pour fonction d'établir le passage d'un objet d'un monde possible à l'autre, du monde des indexeurs à celui des utilisateurs.

Sur quelles bases peut-on assimiler le cercle des indexeurs à un monde possible et celui des utilisateurs à un autre monde possible, ou plutôt à un ensemble de mondes possibles ?

On ne discutera pas les approches linguistiques précédemment présentées : on admettra que la construction d'un monde possible repose sur une formulation linguistique. Mais nous focaliserons, quant à nous, notre attention sur le rôle de l'objet dans la construction du monde possible.

A - Univers des documents et création d'un monde possible

Si l'on considère que l'ensemble des productions éditoriales constitue, pour les indexeurs, le « monde réel », on pourrait penser que l'univers des documents n'est qu'une portion, qu'une partie, de ce monde réel, mais qu'il n'en propose pas, à proprement parler, un autre « état », pour reprendre les termes de Kripke. Or, nous avons tenté de montrer que la sélection des sources en indexation peut se comprendre comme une décontextualisation qui pouvait permettre à une source de se voir attribuée d'autres propriétés, notamment celles de nouveaux usages. La création de l'univers des documents en indexation nous paraît en cela pouvoir être assimilée à la création d'un monde possible : construit à partir des objets du « monde réel » (les sources), le monde possible des documents attribue de nouvelles propriétés à ces objets. L'attribution de ces nouvelles propriétés ne se fait pas dans le cadre d'un discours compris au sens linguistique du terme comme production d'énoncés. La notion de discours documentaire que nous avons proposée se rattache plutôt à la notion de « transformation discursive » de Foucault, qui modifie la situation d'énonciation plutôt que l'énoncé lui-même. En cela, notre conception de la notion de monde possible peut sembler s'éloigner à la fois de celle de Kripke et des reformulations linguistiques dont elle a pu faire l'objet. Néanmoins, il nous semble que nous gardons de Kripke l'idée que la construction d'un monde possible s'établit sur la base d'un objet du monde réel dont on se demande quelles autres propriétés (ici d'usage) il va pouvoir se voir attribué. Par ailleurs, si l'indexation ne produit pas de discours à proprement parler, les descripteurs peuvent être considérés comme l'ancrage linguistique d'un monde possible.

B - Les mondes possibles des utilisateurs

Là encore, notre utilisation de la notion kripkéenne de monde possible repose essentiellement sur la spécificité de la construction d'un monde contraint par un

objet du monde réel. Dans le cas des utilisateurs, on considérera que le monde réel est le monde d'où ils parlent, dans lequel ils se situent. Une situation de recherche documentaire peut se comprendre comme la construction d'un monde possible si l'on considère qu'une requête est établie à partir d'un objet connu (ici un objet de discours que l'on peut nommer) dont on cherche des propriétés inconnues du monde où l'on est[1]. Une requête comme, par exemple, « les expéditions dans le désert », revient à imaginer qu'il existe un monde où cet objet de discours est doté de propriétés qui ne sont pas celles du monde où l'on se place. En ce sens, l'utilisateur espère, d'une certaine façon, trouver dans un système d'information ou une bibliothèque, un « état » du monde qui n'est pas exactement celui qu'il connaît.

C - Problématique des relations entre mondes possibles en indexation

Si l'on accepte de considérer que le cercle des indexeurs constitue un monde possible et que le cercle des utilisateurs suppose un ensemble d'autres mondes possibles, la question qui se pose à l'indexation est celle de la compatibilité entre ces mondes ou encore le passage d'un monde possible à l'autre.

Sur ce point, Kripke souligne que le passage d'un monde à l'autre ne peut se fonder sur l'identification des propriétés d'un objet. Dans son modèle, deux mondes différents ne peuvent être liés sur la base d'une recherche d'identité ; un monde possible est toujours créé à partir d'un monde posé comme réel. Or, dans l'indexation, on aurait deux mondes construits chacun de leur côté sur la base d'objets différents, ce qui rend *a priori* la relation impossible. Il faut donc supposer que l'un des deux mondes, celui des indexeurs ou celui des utilisateurs, contraint l'autre. Se dégage alors l'alternative suivante : soit c'est le monde possible d'un utilisateur qui contraint l'univers documentaire ; soit c'est l'univers documentaire qui contraint la construction de l'univers référentiel des utilisateurs. Si elles sont symétriquement opposables, ces deux solutions ne mettent pas en cause les mêmes enjeux.

Nous examinerons successivement les deux branches de cette alternative.

C.1 - La première – les mondes possibles des utilisateurs contraignent la création de l'univers de référence documentaire – constitue, nous semble-t-il, le modèle dominant en indexation, pour peu que l'on reformule l'approche classique de l'indexation dans le cadre de la théorie de Kripke. Il consiste à intégrer, au sein d'un seul monde possible – celui des indexeurs – le maximum d'autres mondes possibles, c'est-à-dire, le plus souvent, à prévoir, par les descripteurs, le maximum de possibilités d'interrogation. Cette solution ne prend pas en compte le fait que l'univers des documents est construit et à ce titre doublement contraint : par les sources elles-mêmes et par les choix que sont conduits à faire les indexeurs. À moins de confondre bibliothèque et librairie, le choix d'acquisition en bibliothèque ne se réduit jamais aux propositions d'achats des lecteurs. Pour peu que l'on pose la question du document en indexation, il apparaît que les mondes possibles des

[1] Voir sur ce point Le Guern 1991b, p. 71 : « Le contenu informatif d'un document peut être analysé comme constitué de mises en relations de propriétés avec des entités, objets concrets ou non : un objet de pensée, une abstraction, est un objet logique tout autant que les choses du monde. On peut représenter la démarche de l'utilisateur d'un système d'information comme la détermination de l'objet à propos duquel il désire trouver des renseignements, puis la recherche des propriétés attribuées dans les documents du corpus à cet objet, ce qui passe par le repérage des documents qui font mention de cet objet ».

utilisateurs (les « besoins d'information ») ne peuvent être pris en compte au niveau de l'univers des indexeurs lui-même.

C.2 - S'il y a nécessité de rapport contraint entre monde des indexeurs et monde des utilisateurs, la contrainte ne peut s'exercer que dans un seul sens : c'est le monde des indexeurs qui contraint celui des utilisateurs. Dès lors, l'enjeu du descripteur ne se pose plus de la même façon, et d'autres dimensions, que celle des mots, doivent être mises en œuvre. L'opération d'indexation doit disposer de quoi guider l'interprétation, de quoi contraindre la lecture. On dira qu'à ce titre elle met en place une « stratégie d'exposition » des documents. Sur ce point, on rejoint la conception de monde possible proposée par Eco : l'utilisateur se trouve dans la position d'un lecteur qui progresse dans sa recherche d'informations par formulations successives d'hypothèses qui lui sont proposées par la collection documentaire, alors perçue comme « monde réel ». Dans ce cadre, le descripteur n'a plus pour rôle de « prévenir » la formulation linguistique des requêtes d'un utilisateur. L'enjeu ne consiste plus à anticiper la variation linguistique mais à l'exploiter, au contraire.

En ce sens, une approche des fondements théoriques de l'indexation peut s'inscrire dans un programme de recherche tel que le conçoit Dubois : « Prendre la mesure des variations en langue et langages comme une donnée de fait, sinon comme l'essence même des phénomènes de langue [...] ne pas traiter la diversité des interprétations en termes de décalage par rapport à une vérité sémantique du texte mais par rapport à une norme d'interprétation historiquement et socialement située.[1] »

L'enjeu de l'indexation revient donc à exposer comme telle la « norme d'interprétation » des documents qu'elle propose, norme dans laquelle les utilisateurs pourront alors se situer.

Le recours au modèle des mondes possibles établi par Kripke nous a paru indispensable pour poser les enjeux de l'indexation relatifs à la communication entre indexeurs et utilisateurs. Il ne permet pas, bien sûr, d'envisager l'intégralité des aspects de cette relation. Néanmoins, par le modèle de Kripke, le rapport entre indexeurs et utilisateurs peut se percevoir sous l'angle de la problématique précise de la compatibilité entre univers de référence : Kripke établit une notion de monde possible qui exige un ancrage dans un monde réel. Cette conception nous paraît précieuse notamment par la reformulation de la question de l'identité d'un objet entre mondes qui la sous-tend : si l'on peut parler d'un même objet, d'un monde à l'autre, c'est parce que nous-mêmes l'aurons ainsi décidé, dit en substance Kripke. Cette proposition souligne la nécessité en indexation de forcer les différents mondes à se rencontrer, sans pouvoir s'en remettre aux choses telles qu'elles se donnent.

III.2 - Éléments pour une stratégie d'exposition des documents en indexation

En proposant de concevoir l'indexation à travers la notion de stratégie d'exposition des documents, nous essayons de déplacer la problématique classique de l'indexation. Il ne s'agit plus de poser l'existence *a priori* d'objets semblables dans différents mondes (celui des auteurs, des indexeurs, des utilisateurs) et de trouver un mode de désignation commun à tous ces mondes ; il s'agit, au contraire, de

[1] Dubois 1995, p. 93.

partir des objets tels qu'ils se donnent, dans leur singularité, et de définir les moyens qui pourraient permettre de contraindre le regard porté sur ces objets.

Nous entendons, par stratégie d'exposition, l'établissement d'un cadre au sein duquel les utilisateurs peuvent construire leur parcours interprétatif : le cadre d'interprétation fonctionne comme une contrainte interprétative en indiquant dans quel « domaine » ou sous quel « angle » un ensemble de documents peut être perçu.

Quels sont les moyens que l'indexation peut mettre en œuvre pour disposer un cadre qui contraigne l'interprétation des documents ?

On trouve, dans la pratique documentaire classique, les prémices d'une stratégie d'exposition des documents qui recourt aux langages documentaires, et, plus particulièrement, aux langages classificatoires (III.2.1). Ce type de stratégie conduit à des limites que nous avons déjà signalées[1]. Un autre type de stratégie pourrait être envisagé qui s'inspire de la démarche adoptée par les vulgarisateurs scientifiques (III.2.2).

III.2.1 - LE DISCOURS CLASSIFICATOIRE

Les stratégies d'exposition des documents s'appuient, dans les pratiques courantes, sur l'utilisation de classifications documentaires. Après avoir montré en quoi celles-ci pouvaient se comprendre comme des « stratégies », nous examinerons deux aspects du discours classificatoire : les classifications hiérarchiques et les classifications à facettes.

A - Les classifications comme « stratégies d'énonciation de l'offre documentaire[2] »

L'utilisation des classifications comme modes d'exposition des documents est particulièrement visible dans les bibliothèques qui ont opté pour la mise en accès direct des documents[3]. Comme précédemment la notion de politique d'acquisition, le libre-accès des documents constitue une préoccupation récente en France[4], qui fait, depuis ces quinze dernières années, l'objet de débats : faut-il maintenir la classification Dewey pour présenter les documents au public ou faut-il créer de nouveaux modes d'exposition[5] ?

L'introduction du libre-accès en bibliothèque a élargi la problématique de l'exposition à d'autres dimensions que celle du seul langage documentaire. En effet, il est très vite apparu que la logique de la classification Dewey, « unidimensionnelle », ne pouvait apporter toutes les réponses nécessaires à un aménagement de l'espace, lui « tridimensionnel » : dès lors « la classification elle-

---

[1] Voir § II.2.2 dans ce chapitre.
[2] L'expression est d'É. Véron 1990.
[3] Les principes sont les mêmes, la visibilité en moins, dans les systèmes d'information.
[4] Les débuts du libre-accès se situent après la Seconde Guerre mondiale, un développement plus massif se note dans les années 1970 ; toutes les bibliothèques ne proposent pas, à l'heure actuelle, de collections en libre-accès.
[5] Certaines bibliothèques proposent des modes d'exposition des documents par « centres d'intérêt », qui réorganisent les classes de la Dewey : il s'agit là d'un choix « orienté utilisateurs » (« Le centre d'intérêt est un espace logique et matériel dont le véritable centre est le lecteur », Véron 1990, p. 88). Une structuration par centres d'intérêt peut fournir des catégorisations de type : « pays, paysages, voyages » ; « le monde des spectacles » ; « l'art et les artistes » ; « vécu » ; « le temps libre et vos loisirs », etc. (cité *in* Véron 1990, p. 85).

même ne comportant aucune règle d'étalement spatial, un même fonds peut être spatialisé d'une multitude de manières différentes[1] ».

En examinant la mise en espace des collections dans quatre bibliothèques publiques, Éliséo Véron [1990] a pu dégager, à partir de l'utilisation de la classification Dewey, des « discours stratégiques » différents, le plus souvent implicites. La classification Dewey apparaît ici, comme précédemment pour mener une politique d'acquisition, comme un instrument au service d'une stratégie qui ne se dit pas. Mais, ici encore, le moyen finit par prendre plus d'importance que la « fin ».

Au cours de son étude, Véron relève que la diversité des mises en espace n'influe en rien sur les stratégies d'appropriation de la bibliothèque par les lecteurs, comme si la classification constituait déjà, en elle-même, une mise en espace, dont la « qualité » intrinsèque influe peu sur le caractère opératoire : « Nous sommes tentés de conclure à ce propos que le rapport des usagers à la classification implique tout simplement la nécessité d'un système de repérage permettant la constitution d'une stratégie. La classification est une garantie de l'existence d'une convention particulière, de l'absence d'arbitraire. Autrement dit, une classification est indispensable, ne serait-ce que comme élément contre lequel organiser une stratégie, mais cette classification ne nécessite pas des perfectionnements particuliers, et l'on peut soupçonner que n'importe quelle classification, pourvu qu'elle soit stable et régulière, fasse l'affaire.[2] »

Il nous semble que ce propos de Véron souligne le fait qu'une classification fonctionne déjà, en soi, comme une stratégie d'exposition. En effet, ce que ne considère pas toujours Véron dans son étude, c'est que la classification, si elle est utilisée par les professionnels des bibliothèques dans le cadre de stratégie plus ou moins explicite, est avant tout construite par les professionnels eux-mêmes et constitue à ce titre un premier niveau de « discours », le « discours classificatoire ». Dès lors, les observations que Véron formule sur l'efficace de la seule classification peuvent être vues, moins comme un échec des différentes tentatives de mises en espace des collections, que comme un signe de la nécessité absolue de disposer d'un cadre dans lequel les utilisateurs puissent construire leur parcours, leur système d'interprétation.

Véron semble dire que tous les cadres se valent, même si sa préférence va aux cadres les plus « décalés[3] ». La question nous semble plus ouverte, ne serait-ce que parce que les différents types de classifications eux-mêmes n'établissent pas exactement le même type de « discours », de stratégie d'exposition.

---

[1] Véron 1990, p. 11.
[2] *Ibid.*, p. 81.
[3] *Ibid.*, p. 86 : « Les résultats de cette recherche nous ont amenés à une conclusion : le système de classification lui-même est moins important que ne le laissent supposer les discussions passionnées autour de la Dewey. [...] L'enjeu véritable de l'opposition entre les partisans de la classification Dewey (plus ou moins "adaptée") et les partisans de la philosophie des "centres d'intérêt", est celui de choisir entre une classification marquée par une conception du monde et des savoirs qui datent du XIX^{e} siècle, et une classification qui sera inévitablement marquée par une autre idéologie plus actuelle. À cet égard, on peut se demander s'il n'est pas préférable de faire appel à une classification en décalage avec le monde contemporain (comme la Dewey) plutôt qu'à un système qui ne fait que reproduire (cette fois dans l'espace des bibliothèques municipales) la "grille" consacrée autour de nous ».

Dans les deux paragraphes suivants, nous essayerons de montrer comment se manifeste cette diversité du « discours classificatoire ».

B - Les classifications hiérarchiques

On reprend de Bourion et Malrieu [1994] leur approche du plan de classification en termes de « discours classificatoire ». Elles définissent le plan de classification comme un « texte techno-scientifique utilisant une terminologie consensuelle, celle des langues de spécialité qui relèvent de la discipline traitée[1] ». Elles s'interrogent sur la façon dont une classification gère les objets particuliers que sont les textes et posent l'hypothèse qu'un plan de classement, s'il s'apparente au modèle canonique des classifications (logique de type inclusion de classes), repose en fait sur une sémantique de relations tout à fait particulière, qui spécifie un type de discours particulier, le discours classificatoire : « [Un plan de classification] se présente comme une arborescence qui évoque une structure logique canonique de type inclusion de classes, mais la nature ontologique des objets classés ("objets sémiotiques" et non "objets du monde physique") va de pair avec des dimensions de description obligatoirement hétérogènes, en contradiction avec la structure arborescente taxinomique.[2] »

L'étude que mènent Bourion et Malrieu [1994] du plan de classement utilisé par le centre de documentation du CNRS dans le domaine de la psychologie montre que, sous l'adoption explicite d'un cadre canonique et objectif, se déploie un discours classificatoire spécifique aux indexeurs :

- les indexeurs créent des classes ou des sous-classes qui ne se réfèrent pas à des champs de spécialités reconnus par les experts d'un domaine mais qui permettent de ranger des documents qui se situent aux confluents de plusieurs problématiques ; Bourion et Malrieu dégagent sur ce point une sémantique de relations spécifique aux indexeurs, qui ne doit rien à la logique d'inclusion de classes[3] ;

- s'ils modifient la modélisation scientifique d'un domaine, les indexeurs ne se sentent pas toujours tenus d'introduire dans un plan de classement de nouvelles dénominations de classe ; il se passe souvent plusieurs années avant qu'une nouvelle dénomination scientifique soit intégrée dans un plan de classement[4].

Un certain nombre de modifications et de choix sont donc effectués par les indexeurs sur la base d'un référentiel scientifique : c'est ce qui constitue le « discours classificatoire ». Ces changements imposés, insistent Bourion et Malrieu, par la spécificité sémiotique des objets que manipulent les documentalistes, restent cependant peu explicités et problématisés, notamment le rôle implicite attribué au graphisme dans un plan de classement ne fait l'objet d'aucune verbalisation[5]. Nous dirons qu'une classification constitue en cela une stratégie d'exposition implicite des documents, stratégie pas toujours maîtrisée comme telle, empruntant ses

[1] Bourion et Malrieu 1994, p. 84. Nous ne rentrons pas ici dans le détail de leur étude.
[2] *Id.*
[3] *Ibid.*, p. 93-104.
[4] Bourion et Malrieu donnent l'exemple du terme « victimologie », repéré dans la littérature dès 1981, mais définitivement introduit dans le plan de classement en 1983. *Ibid.*, p. 87.
[5] *Ibid.*, p. 116.

moyens non seulement à la langue mais aussi à d'autres systèmes sémiotiques (comme le graphisme).

La particularité du discours classificatoire, un discours « non dit », n'est pas sans incidence sur l'interprétation que peuvent faire les utilisateurs des dénominations des classes d'un plan de classement : les termes spécialisés utilisés dans la classification perdent peu à peu tout lien avec le domaine d'où ils viennent sans que l'utilisateur ne dispose de nouveaux cadres explicites d'interprétation[1]. Faut-il alors « intégrer dans les graphes conceptuels les contraintes liées au genre et au domaine », comme le proposent les deux auteurs à la toute fin de leur étude[2] ? Il ne nous semble pas que ce soit au niveau du langage documentaire que la contextualisation des termes soit la plus efficace. En effet, il nous semble que l'on rencontre alors le problème initial soulevé par Bourion et Malrieu concernant la spécificité sémiotique des objets de l'indexation : on ne peut classer un document comme on classe des objets, des mots, des concepts. Si les termes scientifiques peuvent être classés dans un domaine (une terminologie), parce que ce sont finalement, dans ce cas, des concepts scientifiques (des « choses » pourrait-on dire) qui sont classés, il ne semble pas que des documents, compte tenu de leur spécificité « textuelle », puissent être, au même titre que des termes, classés dans un domaine. Une autre approche devrait pouvoir être dégagée pour tenir compte de la spécificité des textes ; nous en proposons des aspects en II.2.2.

Dans le discours classificatoire fondé sur un langage documentaire, on se heurte à nouveau à deux types de problèmes que l'on a pu régulièrement identifier au cours de cette étude :

- d'une part, le rapprochement explicite de textes se fait *a posteriori* lorsque les textes ont déjà été sélectionnés et distribués en grandes disciplines : stratégie d'exploration des sources et stratégie d'exposition des documents ne sont pas considérées dans leur spécificité ;

- d'autre part, et compte tenu du fait que l'on ne dispose d'aucun *a priori* non formel pour dicter le rapprochement des documents, on en vient à être contraint de réduire un texte non seulement à un mot (indice de classification) mais aussi à un domaine.

Si les classifications hiérarchiques peuvent, en indexation, indiquer la présence d'un discours classificatoire qui tente d'établir une « communication » sur la base d'un référentiel scientifique, elles ne peuvent à proprement parler constituer une stratégie d'exposition des documents qui permette une libre circulation entre textes : la contrainte du langage documentaire comme l'approche en termes de contenu y restent trop fortes.

Une variante du discours classificatoire, articulée sur la notion de « facettes », pourrait-elle être plus efficacement utilisée dans une stratégie d'exposition des documents ?

---

[1] Bourion et Malrieu 1994, p. 117.
[2] *Id.*

C - Les classifications à facettes

La notion de facette, élaborée par Ranganathan[1] pour optimiser le classement des ouvrages en bibliothèque[2], peut s'approcher en termes de « point de vue » sur un objet.

La spécificité des langages documentaires à facettes tient, d'une part, à ce qu'un document y est systématiquement envisagé sous plusieurs angles, et, d'autre part, à ce que les facettes relèvent d'un type de propriétés particulières d'un objet : les facettes s'apparentent en effet aux propriétés « sortales » et non individuelles d'un objet[3]. La notion de facette vise à introduire, dans les langages documentaires, une autre approche que celle du seul contenu : c'est à partir de propriétés extra-linguistiques qu'un objet est classé.

Sans rentrer dans le détail des facettes proposées par Ranganathan[4], on peut donner un exemple[5] de ce que tend à proposer une approche par facettes :

- par la facette Personnalité, le terme « rose » pourra être appréhendé sous l'angle de « l'espèce végétale » ;
- par la facette Matière, le terme « instrument » pourra être perçu sous l'angle de la « musique » ;
- par la facette Énergie, le terme « labourage » sera perçu sous le point de vue « agriculture » ;
- les facettes Espace et Temps reprennent les coordonnées spatio-temporelles indiquées dans un document.

On remarque que le principe de l'analyse par facettes rejoint celui de l'inscription d'un terme dans un domaine, mais la différence tient à ce que plusieurs domaines peuvent être convoqués pour le classement d'un même document[6], même si l'analyse par facettes suppose toujours un rattachement préalable à une classe principale[7].

Ainsi l'indice suivant construit à partir de la *Colon Classification* : « NA 561, J 37, 67 : 8 » devra-t-il se lire « maquette de tour d'un château Tudor » avec, pour cadre

---

[1] Mais dont, comme le suggère Michel Le Guern, on peut identifier les prémices chez Bernard Lamy. Voir Mustafa-Elhadi [1989, p. 180 et suiv.] qui étudie cette filiation historique.

[2] Directeur de la bibliothèque de l'Université de Delhi puis organisateur du réseau des bibliothèques en Inde, Ranganathan crée en 1933 le premier langage documentaire à facettes, la *Colon Classification*, voir Ranganathan 1976.

[3] Ranganathan reprend d'Aristote cinq catégories susceptibles d'appréhender un même objet sous différents points de vue ; ces catégories ont été traduites par les termes français suivants : Personnalité, Matière, Énergie, Espace, Temps ; [Ranganathan 1976].

[4] La littérature sur le sujet est abondante. On peut se reporter par exemple à Vickery 1963 qui propose les plus importantes reformulations des principes de Ranganathan, à de Grolier 1962, 1970 et 1988 pour une mise en perspective, à Salvan 1972 pour une approche comparative, à Mustafa-Elhadi 1989 pour une approche historique et pour une reformulation des concepts en terminologie.

[5] Repris de Salvan 1972, p. 32-33.

[6] Le classement d'un document s'effectue sur la base de l'analyse de son titre, plus ou moins enrichi par la lecture de certaines parties du document et plus ou moins réécrit, Ranganathan 1976.

[7] Pour exemple, on ne donne ici que le premier niveau des classes principales de la *Colon Classification* : « Généralités ; Sciences mathématiques ; Sciences physiques ; Expérience spirituelle, mysticisme ; Humanités et sciences sociales ».

d'interprétation, la classe « beaux-arts ». Chaque élément de l'indice est pourvu d'une signification autonome qui doit permettre une combinaison de sujets. Ainsi « NA 561 », qui se lit « Angleterre, Architecture », devra-t-il permettre de retrouver le document sur la « maquette de tour d'un château Tudor » ; de même de la suite « NA 561, J 37, 6 », qui se lit « Toit, château, Tudor (période) ». Autrement dit, l'indice de base génère plusieurs indices (par « découpage » successif) qui constituent autant d'accès, pour un utilisateur, au même document, selon la facette envisagée (matériau : toit ; temps : Tudor, etc.).

Le principe des facettes introduit par Ranganathan revient donc à restituer, pour chacun des mots retenus dans la définition du sujet d'un document, un contexte d'interprétation de type référentiel, mais là encore ce sont plus les interprétations d'un document que le document lui-même qui sont classées, sans que le point de vue de l'interprétation (la stratégie de lecture adoptée) ne soit, lui-même, spécifié.

Reste qu'en matière de stratégie d'exposition, la notion de facette représente un niveau supplémentaire de transparence par rapport aux précédentes classifications hiérarchiques : elle explicite le point de vue sous lequel il convient de lire un terme, sans laisser à l'utilisateur le soin de se perdre dans le parcours d'une structure arborescente qui répond à la seule logique des indexeurs.

Sans doute une stratégie d'exposition des documents qui reposerait sur le seul usage d'un langage documentaire aurait-elle intérêt à utiliser le principe des facettes, ne serait-ce que parce que l'inscription des termes d'indexation dans un domaine d'interprétation y est systématiquement explicitée. Une étude précise reste à faire sur ce point, qui devrait cependant tenir compte du contexte qui, à l'époque, a présidé à l'instauration du principe des facettes, de façon à écarter ce qui désormais peut se formuler autrement tout en maintenant, parallèlement, les principes de Ranganathan qui restent valides. En effet, à la lumière de ce que nous connaissons désormais de la construction des unités terminologiques, il semble que l'on peut reformuler fructueusement l'approche de Ranganathan, souvent bien complexe dans son système de notation et trop ad hoc dans la définition des facettes.

En effet, comme le rappelle Vickery, « les raisons pour lesquelles l'ancienne forme de classification énumérative, symbolisée par l'arbre de la science, est actuellement périmée sont assez claires. Les sujets hautement spécifiques qui font aujourd'hui l'objet du catalogage sont des sujets composés ; ils ne peuvent être correctement exprimés que par des vedettes-matières combinant deux termes ou plus de deux termes. Chaque terme peut être utilisé dans une grande variété de combinaisons. Aussi est-il nécessaire d'assurer à cette formation une flexibilité totale. C'est la seule manière d'assurer une référence spécifique aux innombrables sujets particuliers. D'autre part, les utilisateurs exigent également de pouvoir, dans une série donnée, disposer de possibilités exhaustives pour conduire une recherche générique ; ils veulent pouvoir repérer un document sur un sujet complexe spécifique non seulement lorsque ce sujet particulier fait précisément l'objet de leur recherche mais aussi lorsque cette recherche porte sur n'importe quel terme collectif incluant l'un des termes de la composition. Ceci implique non seulement que les termes se prêtent à des combinaisons illimitées mais aussi que les relations

génériques (ou relations de classes) soient incluses dans la structure du système. La classification « à facettes » est un moyen d'atteindre ces résultats.[1] »

Par cette longue citation de Vickery, nous voulions indiquer que si, dans les années 30 comme dans les années 60, la classification à facettes pouvait apparaître comme l'un des rares moyens (voire le seul) permettant d'exprimer des sujets « composés », les connaissances actuelles issues de la terminologie peuvent nous permettre d'apporter un autre type de réponse au problème qu'exprime ici Vickery. En effet, la terminologie a pu se livrer à l'étude d'emboîtement de termes (on dit, par exemple, que le terme « base de données » est inclus dans le terme « système de gestion de bases de données ») ainsi qu'à l'étude du rôle classificatoire des « têtes » des unités terminologiques. Sans doute pourra-t-on trouver là des pistes qui permettent de repenser la définition initiale des systèmes à facettes. Nous creuserons certaines d'entre elles dans le chapitre V, mais pas sans avoir préalablement observé le rôle des terminologies dans le discours vulgarisant.

III.2.2 - LE DISCOURS VULGARISANT

Précédemment[2], nous avons émis l'hypothèse que la vulgarisation scientifique et l'indexation présentaient une problématique similaire : celle du passage d'un seul monde possible, où la référence des termes est fixe, à une pluralité de mondes possibles, où la référence des termes est variable. C'est à ce titre que nous pensons que l'examen des principales stratégies discursives à l'œuvre dans le discours de vulgarisation scientifique peut nous aider à formaliser ce que pourrait être une stratégie d'exposition en indexation.

Cette comparaison ne vaut que si les « corpus » sont en vulgarisation comme en indexation établis sur le même principe. En vulgarisation, l'analyste établit un corpus sur la base de la recherche des « formations discursives[3] ». En indexation, nous avons vu qu'il ne pouvait s'agir que d'un idéal théorique dont on voit encore mal, pour le moment, une application généralisée. Reste que nous prendrons comme hypothèse, dans ce paragraphe, que les documents sont ici regroupés dans un « système-archive ». Tenir cette hypothèse permet de maintenir distinctes stratégie d'exploration des sources et stratégie d'exposition des documents.

Cette hypothèse posée, on peut examiner comment le discours vulgarisant procède, de façon à dégager quelques éléments pour élaborer une stratégie d'exposition propre à l'indexation.

A - Stratégie d'exposition en vulgarisation scientifique

Comme s'attache à le montrer Jacobi, la diffusion de connaissance ne se fait pas sans le recours à une stratégie d'exposition[4]. La science des chercheurs ne se laisse pas naturellement « exposer » ; il faut qu'elle soit « mise en scène », ce qui suppose un traitement particulier des terminologies[5].

---

[1] Vickery 1963, p. 4.
[2] Voir § I.3 dans ce chapitre.
[3] *Supra*, § I.4.
[4] Jacobi 1987.
[5] Mortureux 1988, p. 124.

En effet, comme nous l'avons précédemment relevé, le discours de vulgarisation scientifique ne peut, au risque de se nier, ignorer les termes des terminologies spécialisées. Or ceux-ci n'existent et ne font sens que dans le cadre d'un cercle restreint de locuteurs : sortis du monde où ils sont « possibles », ils perdent une grande partie de leur efficience[1]. Le problème consiste donc à maintenir, dans le cadre d'un discours de vulgarisation à large spectre, la présence des termes sans pour autant laisser se dissoudre leur référence.

L'essentiel du travail de vulgarisation consiste alors en une série de reformulations engagées autour des termes, considérés comme des « termes-pivots » : « les mots savants représentent en quelque sorte les traceurs d'une activité de reformulation en train de se faire[2] ».

Deux principaux types de procédé linguistique sont utilisés – les procédés définitionnels et les procédés désignationnels – qui opèrent, tous les deux sur la référence :

- le paradigme définitionnel consiste à établir une « co-occurrence dans un champ discursif donné de plusieurs définitions, paraphrases, gloses, non identiques (formellement, et parfois même sémantiquement)[3] » ;
- le paradigme désignationnel s'apparente au « déploiement d'un ensemble de désignations qui sont approximativement co-référentielles[4] ».

Il nous semble que c'est particulièrement le principe de la constitution de paradigmes désignationnels qui pourrait être exploité en indexation.

En vulgarisation, le paradigme est organisé par le discours qui dispose, autour d'un terme-pivot, d'une série de désignations qui, en contexte, constituent des équivalents approximatifs du terme scientifique visé. Ce que montre la vulgarisation, c'est qu'un mot ne remplace pas un autre mot, mais qu'au contraire les saisies successives que permettent des désignations multiples sont les seules façons de s'approprier la « connaissance » ; comme le note Mortureux, « les locuteurs peuvent jouer du vocabulaire pour désigner une seule et même réalité [...] sans que cette diversité de vocables semble gêner l'identification du référent[5] ».

Sur un plan linguistique, un paradigme désignationnel est constitué d'un ensemble de syntagmes nominaux fonctionnant, dans un corpus donné, en co-référence avec un syntagme nominal pourvu, lui, dans les discours spécialisés, d'une définition explicite[6]. Il y a donc nécessairement présence, dans le discours de vulgarisation, de textes de nature différente, en majorité des textes de nature scientifique. Les « appels » d'un texte à l'autre, parfois explicites (mention de sources ou citation d'auteurs), le plus souvent implicites, se font sur la base des termes spécialisés : en cela, le discours de vulgarisation procède par « focalisations » sur des termes-pivots.

---

[1] Jacobi 1993, p. 74 : « Si personne ne discute l'efficacité des termes spécialisés au sein de petites communautés sociolinguistiques d'experts qui les mobilisent, on sait que, dans une perspective de socio-diffusion, à destination d'un plus grand nombre d'interlocuteurs, non-spécialistes ou novices, ces terminologies cessent d'apparaître comme un excellent vecteur communicationnel pour se muer en obstacle ».
[2] Jacobi 1987, p. 64.
[3] Mortureux 1993, p. 3.
[4] *Id.*
[5] Mortureux et Petit 1989, p. 47.
[6] *Ibid.*, p. 44.

Il n'est pas le lieu de décrire ici tout le « répertoire métalinguistique[1] » dans lequel peut piocher le vulgarisateur. Il nous importe surtout d'illustrer, à ce stade, la possibilité que des procédés linguistiques de reformulation, eux-mêmes contraints par la langue[2], contraignent la lecture et orientent l'interprétation, celles-ci étant guidées par les unités lexicales particulières que sont les « termes ».

L'exemple du discours de vulgarisation scientifique dessine, à gros traits, l'une des stratégies d'exposition possibles en indexation.

B - Éléments du discours de vulgarisation dans la stratégie d'exposition des documents en indexation

Rappelons que l'enjeu, en indexation, d'une stratégie d'exposition est le suivant : l'indexation doit pouvoir contraindre l'interprétation que les locuteurs feront des documents. Pour cela, il faut les leur présenter dans un cadre, un contexte d'interprétation. Ce cadre peut être une classification hiérarchique ou – mieux – une classification à facettes. Ce cadre nous semble pouvoir être aussi celui des terminologies spécialisées, du moins tel que le discours de vulgarisation scientifique en fait usage. Le terme de spécialité nous semble pouvoir fonctionner, en indexation, comme il est utilisé en discours par les vulgarisateurs : il doit permettre de rattacher un document à d'autres textes sans que lui-même ne constitue une unité d'information, de même que le terme ne constitue pas une « unité de connaissance » en vulgarisation mais permet de faire référence aux domaines dont il est question. La mise en œuvre du principe d'exposition des documents inspiré du discours vulgarisant pourrait être la suivante.

Les documents peuvent être classés, dans plusieurs classes le cas échéant, en fonction des terminologies qu'ils mettent en œuvre. Le repérage des terminologies, s'il peut se faire de façon automatisée, n'en demande pas moins un travail de validation manuelle : on a pu précédemment montrer en I.2 que les terminologies n'existaient pas en soi mais qu'elles prenaient sens au travers de textes et de locuteurs. Le principe consiste, comme dans les approches précédentes, à regrouper les textes par « domaines » mais le regroupement s'effectue ici sur la base des termes issus des textes eux-mêmes, et non sur la base d'une analyse de contenu à proprement parler, et ne vise pas à un classement unique, mais au contraire multiple. Un texte peut très bien, comme le montre le discours de vulgarisation scientifique, relever de plusieurs domaines différents à la fois. Les termes utilisés pour classer un document dans plusieurs domaines ne sont pas, en outre, nécessairement ceux qui seront proposés comme descripteurs ; nous y reviendrons.

Beaucoup d'aspects restent à spécifier pour que ce type de stratégie d'exposition des documents puisse réellement voir le jour. Cependant des tentatives de classification de documents sur la base de leurs terminologies ont déjà été expérimentées[3].

Mais là encore, il nous semble important de ne pas confondre les termes utilisés pour la répartition des documents en domaines (au sens terminologique) avec les

---

[1] Jacobi 1993, p. 69-83.

[2] Cf. Mortureux 1993, p. 6 : « Il s'agit là d'une liberté *stratégique* au moins *partiellement déterminée par la structure du lexique* » (c'est nous qui soulignons). Mortureux montre comment l'activité professionnelle de reformulation linguistique faite par le vulgarisateur est linguistiquement contrainte.

[3] Voir, sur ce point, le colloque SFBA (Société française de bibliométrie appliquée) de 1995.

termes utilisés pour circuler entre ces domaines (les descripteurs proprement dits). Les domaines ne sont constitués que pour donner à voir les documents sous un certain angle mais ce n'est pas uniquement sous cet angle-là que peuvent être lus les documents.

Le type de stratégie d'exposition des documents que nous proposons reprend, à rebours, la méthode mise en œuvre par Foucault pour dégager l'espace des pratiques discursives[1] : on commence par considérer les objets de discours dans le cadre des regroupements académiques dans lesquels ils se donnent pour ensuite élaborer des chemins de traverse entre ces différents regroupements.

C'est ici, comme dans le discours de vulgarisation scientifique, la notion de parcours référentiel effectué par le lecteur qui est mise en avant : « La référenciation adéquate peut être vue comme un processus de construction d'un chemin liant différentes dénominations approximatives qui ne sont pas effacées par le dernier choix. Une conséquence en est que plusieurs tentatives de nomination peuvent être retenues comme adéquates, la correction de l'erreur étant alors utilisée comme une ressource interactionnelle pour invoquer des formulations alternatives.[2] »

À la problématique du mot juste, dominante en indexation, la vulgarisation scientifique, tout comme les études plus générales sur la désignation, conduisent à envisager une autre problématique, où l'interprétation n'est pas « trouvée », encapsulée qu'elle serait dans une seule et unique unité lexicale, mais où l'interprétation se « trouve » par le biais d'une construction des objets de discours qu'il est nécessaire de guider, de contraindre.

En fin de ce chapitre, nous retrouvons les convergences entre les pratiques documentaire, terminologique et vulgarisante qui nous avaient initialement permis d'établir une première approche du discours documentaire. Mais, désormais, le « continuum », pour reprendre le terme de Jacobi, qui peut être établi entre ces trois pratiques de diffusion des connaissances, peut trouver, au-delà du rapprochement analogique, quelques fondements théoriques. En effet, la notion de monde possible établie par Kripke, qui nous a permis de problématiser la notion de stratégie d'exposition des documents (c'est le monde possible des documents qui doit contraindre la construction des mondes possibles des utilisateurs, et non l'inverse), peut aussi être vue comme un mode de représentation des problématiques documentaires qui légitime le rapprochement opéré avec la pratique vulgarisante, et, par son biais, avec la pratique terminologique.

IV - Conclusion du chapitre

Alors que, dans son « premier âge », l'indexation se donnait explicitement sous sa dimension discursive[3], on rencontre aujourd'hui bien plus de difficultés pour la faire apparaître. À maintes reprises dans cette étude, à vouloir capter le discours

[1] Voir précédemment le § II.1.1.
[2] Dubois et Mondada 1995, p. 285.
[3] Comme le rappelle Escarpit [1991, p. 153], aux premiers temps des bibliothèques, c'est l'*incipit* d'un document qui était retenu pour le « représenter ».

documentaire, que ce soit sous l'angle de l'exploration des sources ou sous celui de l'exposition des documents, c'est au langage documentaire que nous nous sommes trouvés confrontés, à un langage documentaire qui apparaît en indexation comme l'outil multifonctions par excellence, permettant tout à la fois de sélectionner des sources, d'attribuer des descripteurs et de présenter des documents.

L'un des enjeux de ce chapitre était précisément de montrer que le recours systématique à un même type d'outil lexical ne devait pas masquer les différents types d'opération que l'indexation réalise sur les discours.

Pour ce faire, nous avons multiplié les pistes pour approcher la notion de discours documentaire, particulièrement absente dans la littérature classique sur l'indexation : chacune des approches proposées ici (approche comparative de pratiques ou approche par reformulation de modèles) pourrait, à elle seule, constituer l'objet d'une étude, dont nous n'avons ici proposé qu'une esquisse.

En dépit du caractère suggestif qui caractérise sur bien des points notre démarche, il nous semble que nous avons pu faire apparaître la nécessité – pour la détermination des descripteurs proprement dite, par laquelle se définit habituellement l'indexation – de disposer d'une perspective plus large, qui englobe la problématique de la sélection des sources comme celle de l'exposition des documents. En effet, c'est en appréhendant l'indexation sous cet angle élargi que se dégagent deux systèmes de contraintes qui pèsent sur l'indexation proprement dite, sur le choix *in fine* des descripteurs :

- un premier système de contraintes est représenté par les sources elles-mêmes. Sur ce point, l'approche de Foucault, quelle que soit la difficulté de son utilisation dans un cadre professionnel particulier, montre qu'il existe des règles – des règles discursives – qui président à l'existence des textes. L'indexation ne peut pas ignorer les conditions d'existence des discours, sous peine de les considérer comme un amas d'objets sans cohérence dans lequel on pourrait piocher impunément. La problématique des politiques d'acquisition nous semble sur ce point pouvoir bénéficier d'un ancrage théorique tel que le modèle de Foucault en offre une forme générale.

  Outre la difficulté de l'entreprise, les conséquences d'un tel choix théorique sur l'approche du descripteur lui-même ne sont pas négligeables : en effet, ce que suggère la démarche de Foucault, comme le met bien en valeur Milner[1], c'est que toute pratique discursive est saisie dans un réseau de termes spécifiques qui constitue ses propres objets de discours. Par conséquent, une pratique d'indexation qui s'inscrirait dans le modèle théorique des formations discursives, s'appliquerait nécessairement à choisir des descripteurs qui soient issus des discours eux-mêmes. L'affectation d'autres mots, issus d'autres « histoires » dans les termes de Foucault, ruinerait complètement l'intérêt de la démarche. Autrement dit, l'indexation ne peut être dans un tel cadre qu'une *extraction de mots de discours* ;

- le second système de contraintes qui pèse sur l'indexation est, lui, lié à la problématique de l'identité des objets en indexation. Dans la première partie de cette étude, nous avons examiné le modèle implicite sur lequel reposait la croyance en la stabilité référentielle des objets en indexation. Dans cette

[1] Milner 1989, p. 66.

> seconde partie, nous avons cherché à redéfinir cette question en la posant dans le cadre théorique des mondes possibles établi par Kripke. Problématisée dans ce cadre, la question de la stabilité référentielle ou encore de l'identité des objets s'exprime clairement en termes de contraintes : qui, des indexeurs ou des utilisateurs, définit les objets à partir desquels vont pouvoir être examinées les différentes propriétés ? Du fait qu'il s'agit de donner à manipuler des objets dont il est question dans les textes d'une collection documentaire, seuls les indexeurs sont à même de déterminer l'ensemble des objets identifiables. De là, la nécessité pour l'indexation de contraindre le parcours interprétatif de l'utilisateur. De là aussi la nécessité de repenser la morphologie du descripteur : pour pouvoir établir une relation entre univers de référence, les descripteurs doivent présenter un certain degré de « rigidité ». En cela, les descripteurs doivent être nécessairement des *termes*.

Dans le cadre d'une approche discursive de l'indexation, la problématique du descripteur se laisse approcher sous une forme sensiblement différente que celle proposée dans les normes ou la littérature classique ; nous spécifierons, dans le prochain chapitre, les deux caractéristiques dégagées ici du descripteur : unité extraite du discours et terme.

À l'issue de ce chapitre, il apparaît que le processus de l'indexation peut être appréhendé sans qu'il soit nécessaire de recourir à la notion de langage documentaire, à la notion d'uniformisation lexicale. La notion de discours documentaire, définie comme espace d'organisation spécifique des documents – spécifique par son mode d'exploration des sources et par son mode d'exposition des documents –, si elle n'a pas été ici entièrement définie, apparaît non pas comme ce qui homogénéise les documents, mais plutôt comme ce qui permet de maintenir, ensemble et accessible, une diversité de documents hétérogènes.

CHAPITRE V
LA PROBLÉMATIQUE DU DESCRIPTEUR

Nous proposons dans ce chapitre une reformulation de la problématique classique du descripteur en indexation[1], qui s'appuie, d'une part, sur le modèle linguistique du lexique et de la référence que nous avons proposé au cours de la première partie de cette étude, et qui repose, d'autre part, sur la notion de discours documentaire que nous avons précédemment établie (chapitre IV).

Cette reformulation paraît nécessaire au vu des remarques que nous avons pu faire :

- l'expression du contenu d'un texte par un ou plusieurs mots relève d'un effet d'interprétation, obtenu par un lecteur sur la base de l'interprétation *en discours* d'une unité lexicale (voir chapitre II). Sur ce point, nous avons proposé de concevoir l'indexation comme une opération exploitant deux propriétés linguistiques des unités lexicales : leur autonomie lexicale, qui repose sur la signification lexicale, que l'on peut approcher par la notion de stéréotype ; leur autonomie référentielle, qui, créée en discours sur la base de la signification lexicale, construit un effet de stabilité référentielle. Toutes les unités lexicales du français ne présentent pas conjointement ces deux propriétés : nous préciserons dans ce chapitre le type d'unité linguistique pouvant être utilisé en indexation. Ce faisant, nous pourrons fonder, du point de vue d'une théorie linguistique, l'emploi, dans toutes les pratiques d'indexation, des unités lexicales de catégorie nominale : le choix de telles formes dépasse la seule « convention d'écriture » ou encore les seules raisons de « commodité pratique[2] ».

- la stabilité référentielle, la stabilité du référent désigné, relève, elle aussi, de l'interprétation (voir chapitre III). Nous avons vu que cet effet interprétatif devait être « contraint ». Le mode de constitution d'une collection documentaire doit permettre de construire des *thèmes de discours*, c'est-à-dire d'établir un parcours interprétatif cohérent : nous avons insisté dans le chapitre IV sur la nécessité de disposer d'un espace de textes « lisible ». Sous cet angle est apparue une autre propriété, de nature « logique » cette fois, que

[1] Présentée au § I du chapitre II.
[2] Voir chapitre I, § II.2, le discours de la norme.

devait révéler une unité lexicale pour être utilisée en indexation : elle doit présenter un certain degré de « rigidité ».

L'objectif de ce chapitre consiste à intégrer, au sein d'une même description, les différentes facettes du descripteur qui nous sont, au cours de cette recherche, apparues essentielles. Le descripteur doit être pourvu :
(i) d'une signification lexicale autonome ;
(ii) d'un pouvoir référentiel ;
(iii) d'une certaine rigidité qui permette de trouver, pour un objet donné, des propriétés que l'on ne lui connaissait pas, la transformation de ces nouvelles propriétés en objet nouveau constituant un thème de discours[1].

C'est l'articulation de ces trois propriétés que nous étudierons dans ce chapitre.

Nous maintenons ici la dénomination « descripteur » pour qualifier le produit de l'indexation bien que l'emploi que nous ferons du terme ne doive plus rien aux discours qui l'ont créé[2]. Il nous paraît important de garder cette dénomination pour deux raisons :

- d'une part, l'approche du descripteur que nous proposerons ici distingue des propriétés qui, nous semble-t-il, caractérisent également les descripteurs des langages documentaires (notamment les propriétés linguistiques d'autonomie lexicale et de pouvoir référentiel) ;

- d'autre part, le terme « descripteur » permet de souligner l'usage documentaire qu'il est fait de certains types d'unité lexicale en indexation. Cette utilisation nous semble caractéristique de la pratique d'indexation. C'est cette spécificité que nous voudrions mettre en valeur.

En cela, ce chapitre proposera une synthèse des différentes hypothèses présentées jusqu'alors. Cette synthèse permet de confronter, au sein d'un même modèle, les approches classiques du descripteur et de l'indexation avec l'approche discursive de ces mêmes objets que nous défendons dans cette recherche.

Pour montrer que ce sont les propriétés des discours eux-mêmes qui contraignent la morphologie[3] du descripteur, nous procéderons de la façon suivante :

- nous commencerons par définir l'enjeu du descripteur comme unité de discours (I). Après avoir montré la difficulté, dans le cadre des problématiques classiques, de penser le descripteur comme unité de discours, nous proposerons un modèle théorique où une telle problématique peut prendre un sens : ce modèle est celui des « chaînes de référence » ;

---

[1] Voir chapitre II § II.1.
[2] Voir la norme Z 47-102 (1978) : « Mot ou groupe de mot retenus dans un thésaurus et choisis parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document ou dans une recherche documentaire », ou encore Chaumier 1978, p. 30 : « Les descripteurs, appelés parfois encore mots-clés, sont les termes qui sont autorisés, à l'exception de tout autre, pour l'indexation des documents et des questions. Ils servent à représenter les concepts ou notions des documents ou des questions. Un descripteur peut être formé d'un mot ou d'une expression ».
[3] Par « morphologie » du descripteur, nous entendons la forme linguistique par laquelle se donne le descripteur.

- nous rentrerons ensuite dans l'examen plus précis de la morphologie du descripteur, en privilégiant d'abord un point de vue logique (II). C'est de ce point de vue que nous discuterons les formes possibles du descripteur : nom propre et/ou description définie ;

- nous resserrerons, pour finir, notre analyse sur la seule forme linguistique du groupe nominal (III) : la morphologie du descripteur sera alors discutée du seul point de vue linguistique. Deux modèles de représentation linguistique seront alors sollicités pour nous permettre de définir le descripteur d'un point de vue formel.

I - Enjeu du descripteur comme unité du discours

La caractéristique du descripteur d'être une unité de discours a clairement été établie par Michel Le Guern[1] et les membres de l'équipe SYDO[2] qui, les premiers, ont dégagé la dimension référentielle de l'indexation et spécifié, sous cet angle, les propriétés logiques et linguistiques que devait présenter le descripteur.

Si l'essentiel des remarques proposées dans cette recherche repose sur cette approche du descripteur, nous avons cherché à reprendre les hypothèses du groupe SYDO sous un angle un peu décalé, privilégiant le traitement des sources documentaires que supposait une telle approche et explorant les dimensions du discours qui nécessairement devaient être en jeu.

C'est également en nous plaçant sous l'angle de l'interdiscursif que nous étudierons l'enjeu du descripteur comme unité de discours. Si nous arrivons aux mêmes conclusions que Michel Le Guern, nous empruntons une voie qui, nous l'espérons, pourra fructueusement compléter son modèle[3].

Avant d'exposer le cadre dans lequel nous étudierons la morphologie du descripteur (I.3), nous préciserons les contours de notre problématique : si le descripteur est nécessairement une unité de discours (I.1), toute unité extraite du discours ne constitue pas, pour autant, un descripteur (I.2).

I.1 - Problématique : le descripteur est nécessairement une unité extraite du discours

Nous rappelons brièvement les fonctions attendues du descripteur, fonctions qui, reformulées dans le cadre d'un modèle linguistique du lexique et de la référence, mettent en évidence ses propriétés discursives. Si une telle approche est théoriquement fondée, elle ne se heurte pas moins à la réalité des pratiques d'indexation qui ne procèdent que très rarement à l'extraction d'unités de discours. Nous réexaminerons sur ce point les discours normatifs concernant aussi bien l'indexation que la recherche documentaires : il apparaît que les définitions qui s'y trouvent entrecroisent des éléments de nature singulièrement hétérogène.

---

[1] Le Guern 1984, 1989, 1991a, 1991b entre autres.
[2] Bouché 1989, Metzger 1988 par exemple.
[3] Présenté ci-après § III.2.1.

I.1.1 - RAPPEL DES FONCTIONS ATTENDUES DU DESCRIPTEUR

Nous reformulons brièvement dans ce paragraphe les deux fonctions du descripteur que nous avions dégagées dans le chapitre II. Nos interrogations initiales étaient les suivantes :

A - Comment le descripteur peut-il remplir sa fonction de « représentation du contenu » d'un document ?

B - Comment le descripteur peut-il fonctionner comme un « accès stabilisé » à une collection documentaire ?

Notre réponse à ces deux questions fait ressortir la nécessité, pour le descripteur, d'être une unité extraite des documents.

A - Fonction 1 du descripteur : « représenter le contenu » d'un document

En prenant en compte la spécificité des objets manipulés en indexation (des textes), on peut montrer que l'« information » trouvée par les utilisateurs ne peut être, d'un point de vue linguistique, que construite *via* l'examen des mentions d'un objet de discours dans une collection de documents.

La problématique de l'indexation devrait donc être celle de la détection des objets de discours. Or la spécificité d'un objet de discours tient à ce qu'il ne correspond pas à une dénomination lexicale, ni même à un ensemble de dénominations lexicales. Comme le montre l'analyse du thème discursif menée par Marandin, l'objet de discours est le fruit d'une lecture qui permet de transformer des énoncés relatifs à plusieurs référents en propriétés interprétables comme étant relatives à un seul référent discursif : ce dernier référent peut être nommé dans le texte lui-même mais peut aussi en être inféré. Par ailleurs, pour un même texte, plusieurs thématisations sont possibles. Il n'est donc pas envisageable d'assimiler les descripteurs à des thèmes de discours, tels que Marandin les définit. Ils doivent plutôt être les différents référents discursifs qui, dans une collection documentaire, permettent de les construire. C'est sous cet angle que nous les aborderons dans ce chapitre. Notons d'emblée que cette approche, outre qu'elle implique que les descripteurs soient nécessairement issus des sources elles-mêmes, ne peut fixer à l'avance le nombre de descripteurs qui seront extraits d'un texte : à première vue, le nombre de descripteurs pourra être supérieur à celui habituellement observé dans les pratiques d'indexation[1].

---

[1] Notons toutefois que les pratiques d'indexation sont très variables sur ce point. L'expérimentation que nous avons menée montre que, pour un même document, l'indexation peut produire de 1 à plus de 20 descripteurs. De nombreux paramètres peuvent expliquer ces divergences, nous en citons quelques-uns sans entrer dans le détail : (i) le type de langage documentaire utilisé (les langages documentaires « synthétiques » de type Rameau permettent une indexation par une seule vedette-matière ; l'utilisation de thésaurus favorise un nombre plus élevé de descripteurs) ; (ii) le type de stockage des descripteurs (le stockage sous forme papier oblige à réduire le nombre de descripteurs ; le stockage sous forme informatique, comme dans les bases de données bibliographiques, permet de multiplier le nombre de descripteurs), etc.

B - Fonction 2 du descripteur : permettre un « accès stabilisé » à une collection documentaire

Si, interprété en discours, un descripteur a pour fonction de permettre la construction d'objet de discours, quelle est sa fonction hors discours, lorsqu'il constitue un accès, décontextualisé, à une collection documentaire ?

Sur ce point aussi apparaît la nécessité, pour le descripteur, d'être un mot issu du discours. En effet, nous avons relevé précédemment que les descripteurs issus d'un langage documentaire (ils sont alors mots de lexique) aboutissaient à des impasses :

- dans le chapitre II, nous avons noté le décalage entre lexique et discours au moment de l'appariement dans un système d'information classique. Le descripteur n'est dans ce cas qu'un mot du lexique, pourvu de sa seule signification lexicale, dans lequel l'utilisateur investit, lui, un objet de discours pourvu de propriétés référentielles, spécifiques à son univers de référence. L'incompatibilité est alors garantie, d'autant que nous avons parallèlement observé, dans le chapitre IV, que les relations établies à l'intérieur d'un langage documentaire ne pouvaient permettre, contrairement à ce que les discours normatifs prétendent, de contraindre l'interprétation référentielle des descripteurs : les relations dans un langage documentaire sont établies de façon *ad hoc* par les indexeurs sur la base d'un ensemble de documents qui n'est plus celui qui est effectivement soumis à l'indexation ;

- dans le chapitre IV, nous avons, par le biais du modèle des mondes possibles, examiné comment pouvait se réaliser le « transport » d'un objet d'un univers de référence à l'autre. Il est apparu nécessaire de concevoir l'indexation comme un lieu de création de la « réalité » (le monde des indexeurs est perçu par les utilisateurs comme le « monde réel ») ; ce n'est qu'à cette condition que deux objets peuvent être tenus pour semblables. De là, la nécessité pour le descripteur de pouvoir d'emblée, en tant qu'accès à ce monde, pouvoir référer. Cependant, hors discours, le descripteur ne peut référer qu'à une classe d'objets, ce n'est qu'en discours qu'il pourra référer à un individu spécifique (un référent discursif). C'est sous l'angle de la dialectique de cette double dénomination (à une classe et à un élément de cette classe) que nous aurons également à examiner le descripteur.

Il ressort que même son emploi hors discours ne peut faire considérer le descripteur comme un mot extrait du lexique.

Les deux fonctions du descripteur reformulées ici (élément de thématisation et accès à une collection documentaire) engagent à le considérer comme étant une unité extraite des textes eux-mêmes.

Cette approche du descripteur implique une redéfinition de l'indexation et de la recherche documentaires, que nous essaierons de situer par rapport aux approches classiques. Cette confrontation fera apparaître l'hétérogénéité des éléments par lesquels sont habituellement définies l'indexation et la recherche documentaires.

I.1.2 - RÉEXAMEN DES APPROCHES NORMATIVES

L'approche du descripteur comme unité extraite du discours fait percevoir les deux processus d'indexation et de recherche documentaires sous des angles qui les éloignent des approches classiques[1] :

- l'indexation y est vue comme une *extraction*, supposant une mise à disposition des textes intégraux des documents au sein même des systèmes d'information documentaires, ne présupposant ni sélection ni traduction de « concepts » et appelant, dans sa version optimale, un traitement automatisé. Les fondements linguistiques de l'indexation trouveraient-ils leur meilleure application par les procédés techniques actuels ? Ou bien avons-nous succombé, dans notre essai de formalisation des principes de l'indexation, aux mirages de la technologie moderne ? Nous essaierons de montrer que, tout comme notre approche est, sans aucun doute, conditionnée par l'environnement technique actuel, la définition normative de l'indexation repose, elle aussi, sur un état historique de la technologie[2] ;

- la recherche documentaire, que nous aborderons ici uniquement pour donner une idée des conséquences qu'implique notre approche du descripteur, se laisse, elle aussi, entièrement revisiter. La contrainte que fait peser le système de recherche documentaire sur l'utilisateur ne s'effectue plus au même niveau. Il ne s'agit plus, pour lui, de choisir parmi un ensemble de mots mais parmi un ensemble de « choses », de référents discursifs. Cet aspect de la recherche documentaire, outre qu'il s'exprime, lui aussi, dans des réalisations informatiques récentes, rejoint des recherches plus contemporaines menées sur les situations de recherche documentaire.

A - L'indexation revisitée

Nous revisiterons le modèle classique de l'indexation en adoptant deux approches :

A1 - Dans la première, l'indexation est appréhendée comme un type de lecture : l'indexation classiquement réalisée par le biais d'un langage documentaire s'oppose à l'indexation-extraction que nous proposons par le type de lecture que respectivement elles supposent ;

A2 - Dans la seconde, l'indexation est appréhendée dans le cadre des typologies courantes : l'indexation classique et l'indexation-extraction laissent apparaître des points de convergence et des points de divergence, qui soulignent le rôle des évolutions technologiques dans les approches de l'indexation.

---

[1] Dans leur acception normative, l'indexation est vue sous le seul angle de l'attribution de descripteurs et la recherche documentaire sous celui de l'appariement de descripteurs (modèle de l'*Information Retrieval*).

[2] Le Guern (communication personnelle) et Dubois [1995] incitent à prendre en compte, dans l'approche théorique de pratiques professionnelles comme l'indexation, le rôle du contexte technologique dans lequel elles ont pu être définies. Cf. par exemple Dubois 1995, p. 93-94 : « Un des moyens de clarifier les questions [celles des contraintes technologiques dans les systèmes d'information] est de resituer les développements technologiques récents (informatisation) dans une histoire des technologies de l'information et de la matérialisation du langage ».

A1 - Types d'indexation et manière de lire

Envisagée sous l'angle du thème discursif proposé par Marandin, l'indexation « classique[1] » consiste à déterminer le thème d'un texte : le descripteur est directement un thème de discours. Dans ce cas, c'est l'indexeur qui détermine le thème d'un texte ; sa lecture le conduit nécessairement à déterminer un seul thème parmi les différents possibles (un thème est une « manière de lire[2] »). L'enjeu de l'indexation se concentre alors sur la question de la transmission de l'interprétation thématique qui a été faite du texte : c'est la difficulté d'une telle transmission qui conduit à la constitution du langage documentaire.

Notre approche de l'indexation consiste, elle, à transférer le lieu de la thématisation, de l'indexeur à l'utilisateur. Ce déplacement appelle une tout autre définition du descripteur. Permettre à un utilisateur de construire son interprétation suppose d'« extraire » d'un texte non des thèmes mais la chaîne, ou les chaînes, de référents discursifs susceptibles de les construire : l'indexation doit donc se situer *en deçà de l'interprétation.*

Peut-on demander à un analyste humain d'indexer un texte sans l'interpréter ? Alors que, sur un plan pratique, une manière fiable d'éviter toute interprétation au niveau de l'indexation lui-même consiste à faire exécuter, de façon automatique, une extraction des référents discursifs, on peut, sur un plan théorique, distinguer deux types de lecture, susceptibles de produire deux types d'indexation différents. L'introduction d'un traitement automatique pour extraire des objets de discours constitue ici une facilité matérielle : elle n'est en rien définitoire de l'approche que nous proposons.

On peut, à la suite de Ricœur[3], distinguer deux « manières de lire[4] » :
- l'attitude explicative qui consiste à « prolonger et renforcer le suspens qui affecte la référence du texte » ;
- l'attitude interprétative qui consiste à « lever le suspens et à achever le texte en parole actuelle ».

Il s'agit donc de déterminer à quelle manière de lire se rattache l'indexation :
- à la manière explicative, l'indexation maintient les différentes potentialités de thématisation et laisse ouverts à l'utilisateur tous les parcours interprétatifs ;
- à la manière interprétative, elle effectue un choix d'interprétation et sélectionne un thème possible parmi plusieurs.

Les descripteurs utilisés dans l'un ou l'autre type d'indexation seront alors nécessairement différents. D'un point de vue logique, il apparaît que :
- l'indexation explicative exploitera principalement les « désignateurs accidentels », susceptibles de construire autant d'objets de discours qu'il y a de « mondes possibles », d'utilisateurs ;
- l'indexation interprétative privilégiera les « désignateurs rigides », dont le nom propre, qui garantit la permanence « thématique », si tant est que la permanence thématique puisse être assimilée à la permanence référentielle[1].

---

[1] Celle qui utilise un langage documentaire.
[2] Marandin 1988, p. 86.
[3] Ricœur 1986, p. 151.
[4] Cette distinction recoupe en partie celle d'Escarpit présentée au chapitre II (lecture objective *versus* lecture projective), Escarpit 1991, p. 129 par exemple.

Il y a donc, d'un point de vue théorique, la possibilité de choisir le type d'indexation que l'on veut réaliser. L'indexation normative est plutôt le fruit d'une lecture interprétative, tandis que nous proposons une approche de l'indexation qui suggère une lecture explicative.

Deux raisons d'ordre différent nous paraissent pouvoir expliquer le choix de l'approche classique en faveur d'une indexation de type interprétatif[2] :

- d'une part, elle correspond à la représentation sous-jacente du sens en indexation[3]. C'est sur la base d'une croyance en un sens vrai et unique du texte que l'indexation interprétative trouve sa légitimité. On remarquera sur ce point que les manuels et traités d'indexation ne masquent pas cette dimension interprétative de l'indexation puisque l'interprétation que l'indexeur fait d'un texte est supposée être la seule possible[4] ;

- d'autre part, elle correspond à un état de la technique spécifique de la fin des années 70 qui a vu fleurir l'essentiel des textes normatifs sur l'indexation. En effet, l'indexation y est définie dans les conditions dans lesquelles elle se pratique dans la majeure partie des cas : c'est-à-dire « manuellement », par une lecture « humaine ». Or la manière de lire explicative, qui se situe en deçà de l'interprétation, reste sans aucun doute difficile à réaliser par un « humain ».

Ces deux contraintes de nature différente conduisant à privilégier un certain type d'indexation méritent donc d'être « désintriquées ». Le modèle de l'indexation défini par les normes ne peut servir de modèle de référence, trop contraint qu'il est à la fois par le type de représentations linguistiques qu'il suppose et par l'état de la technique sur lequel il repose (peu ou pas de systèmes automatisés d'extraction d'objets de discours).

Il apparaît en cela nécessaire de dégager une approche de l'indexation plus théorique qui permette de faire de véritables « choix » professionnels. De ce point de vue, l'approche linguistique du thème discursif par lequel on peut définir des « manières de lire » constitue une des voies possibles. On montre en effet que :

- on peut effectuer une indexation qui soit une interprétation figée et partielle du texte[5], et l'enjeu consiste alors à transmettre cette interprétation ; la morphologie du descripteur reste alors à définir expressément dans ce sens ;

- on peut effectuer une indexation inachevée d'un point de vue interprétatif[6] ; la morphologie du descripteur devra alors être déterminée de ce point de vue :

---

[1] Nous revenons sur ce point au paragraphe suivant II.2.

[2] Nous appellerons désormais « indexation interprétative » ou « indexation de type interprétatif » une indexation issue d'une lecture interprétative des textes (qui détermine un thème). Par opposition, nous appellerons le type d'indexation que nous proposerons « indexation explicative » ou « indexation de type explicatif ».

[3] Voir la première partie de cette étude.

[4] On peut lire sur ce point Neet 1990.

[5] Comme le rappelle Marandin [1988, p. 86], « thématiser, c'est stabiliser un état du monde raconté et se satisfaire d'un monde partiel ».

[6] Ce qui ne signifie pas que l'indexation ainsi comprise fabrique des ambiguïtés interprétatives. Pour un utilisateur, un lecteur donné, un texte n'est pas ambigu ; voir, sur ce point, Marandin, [*Id.*] : « Le sens d'un texte est peut-être inépuisable, sans doute pluriel,

c'est sur celle-ci que nous nous attacherons plus particulièrement dans ce chapitre.

Nous avons essayé de montrer les ressemblances et les différences entre l'indexation-extraction que nous proposons et l'indexation classique. Les deux types d'indexation se laissent analyser par le biais de la même notion de thème discursif et par une approche discursive du descripteur. Cependant, alors que, dans un cas, la thématisation est le fruit de l'indexation, dans l'autre, elle est le résultat d'une recherche d'information. Pour distinguer les deux types d'indexation de ce point de vue, nous avons proposé d'appeler « indexation interprétative » celle qui produit des thèmes de discours et « indexation explicative » celle qui extrait des éléments textuels permettant de construire des thèmes de discours.

Si notre approche de l'indexation-extraction peut être opposée à l'indexation-assignation[1] sur le critère linguistique du mode de lecture, comment situer notre proposition à partir de critères proprement documentaires ?

La littérature professionnelle propose plusieurs typologies de l'indexation, dans lesquelles nous essaierons de nous situer.

A2 - Typologies classiques de l'indexation

Comme le rappelle Le Loarer[2], « il est fréquent, dans le milieu des professionnels de l'information et de la documentation, d'opposer différents types d'indexation ». On parle en effet, comme de variantes d'un processus dont on finit par ne plus savoir en quoi il consiste, d'indexation manuelle ou humaine, d'indexation automatique ou en texte intégral, d'indexation pré- ou post-coordonnée, d'indexation libre ou contrôlée, etc.

La notion d'indexation-extraction ne semble pas être retenue pour figurer parmi cet ensemble de variantes : elle est en effet le plus souvent opposée à celle d'indexation[3].

En reprenant les distinctions habituellement utilisées par les professionnels, on peut néanmoins tenter de situer notre approche. Nous pourrons ainsi discuter les critères retenus dans les typologies classiques en tâchant de resituer le rôle des techniques en indexation :

- opposition 1 : indexation manuelle (ou humaine) *versus* indexation automatique. La présence/absence d'un traitement informatisé distingue clairement les deux approches. Cependant, la nature du traitement automatisé qui les sépare peut être très différente. L'indexation automatique peut

---

mais la compréhension que l'on en a se marque par ce que l'on retient du texte : l'interruption de la construction du monde est constitutive du monde textuel » ; dans ce cadre, une interprétation correspond au « monde textuel » créé par une lecture.

[1] On appellera « indexation-assignation » l'indexation qui recourt à l'utilisation d'un langage documentaire pour assigner à un document des descripteurs. L'indexation-assignation correspond aux pratiques courantes de l'indexation, que l'on a appelé aussi dans cette recherche « indexation classique ».

[2] Le Loarer 1994, p. 158-161. Nous nous inspirons, dans les lignes qui suivent, de la synthèse des typologies de l'indexation qu'il propose.

[3] Chaumier 1989, p. 15 : « Le terme indexation est souvent opposé à celui d'extraction » ; l'auteur précise que, lui, ne partage pas cette position.

consister en la construction d'un index de plusieurs types : documentaire[1], informatique[2], statistique[3], linguistique[4], conceptuel[5], mixte[6], etc.

Au gré des évolutions technologiques, l'indexation automatique emprunte ses méthodes à de multiples domaines de savoir. Le rapport qu'entretiennent l'indexation humaine et l'indexation automatique n'est pas toujours spécifié. Rares sont les systèmes qui reposent sur une simulation de l'indexation « manuelle » ; plus nombreux sont ceux qui visent à produire les mêmes résultats qu'elle. Cependant peu se dotent d'un modèle explicite du descripteur qu'ils prétendent reproduire. Devenue partie prenante dans le domaine des « industries de la langue[7] », l'indexation documentaire, appréhendée sous l'angle automatique, devient plus un terrain d'expérimentation pour les techniques qui traitent automatiquement des textes qu'un objet d'étude à formaliser et/ou à automatiser. En tant que domaine d'application, elle est située sur le même plan que la traduction automatique, l'enseignement assisté par ordinateur, la vérification orthographique, etc. La spécificité de l'indexation ne peut être prise en compte lorsqu'elle est comprise comme domaine d'application des industries de la langue[8]. Autrement dit, les pratiques d'indexation manuelle ne bénéficient que très rarement des domaines de savoir convoqués dans ce type d'indexation automatique.

Sur ce point, notre approche de l'indexation-extraction, si elle appelle un traitement automatisé des textes, se veut également valable dans le cadre d'une pratique manuelle. Si l'opposition indexation manuelle/indexation automatique doit être maintenue, il faut pouvoir fonder cette distinction. Dans la plupart des systèmes automatiques actuels, il n'y a pas de véritable opposition entre indexation manuelle et indexation automatique ; elles évoluent plutôt de façon parallèle, souvent dans une mutuelle ignorance.

- opposition 2 : indexation manuelle (ou humaine) *versus* indexation en texte intégral. Cette opposition repose implicitement sur la précédente, l'indexation en texte intégral étant implicitement assimilée à l'indexation automatique. Le critère central de l'opposition, qui reste le plus souvent en retrait, concerne ici le type de repérage des descripteurs effectué : repérage sélectif ou repérage exhaustif. Ce critère est délicat à manipuler dans la mesure où il tend à confondre lecture sélective du texte (l'indexeur humain, contrairement à l'indexeur machinal, n'a pas à parcourir tout le texte) et repérage sélectif de

---

[1] Fondé sur le repérage, dans un texte, de descripteurs issus d'un langage documentaire.
[2] Au sens que donne la note 2, p. 169.
[3] Voir, par exemple Salton 1986, 1988.
[4] Extraction de groupes nominaux, voir sur ce point les communications présentées au colloque de la SFBA (Société française de bibliométrie appliquée) en 1995.
[5] Voir, par exemple, Collas et Chartron 1994.
[6] Voir, par exemple, les communications présentées au colloque de la SFBA (Société française de bibliométrie appliquée) en 1995.
[7] Terme créé aux débuts des années 80 par analogie avec le terme « industries de l'information » pour désigner « l'ensemble des activités qui visent à faire manipuler, interpréter ou générer par les machines le langage naturel écrit ou parlé par les humains », Carré et *al.* 1991, p. 9.
[8] *Ibid.*, p. 279 : « En fait, ces améliorations [apportées par les industries de la langue] ne sont pas généralement la conséquence d'une meilleure connaissance des phénomènes, mais plutôt d'une meilleure adaptation des produits aux besoins ».

descripteurs (l'indexeur machinal, contrairement à l'indexeur humain, ne sélectionne pas, ne trie pas les descripteurs).

La sélection de descripteurs est vue comme une intelligence du texte propre aux indexeurs, tandis que le repérage exhaustif des descripteurs est vu comme une « faiblesse » du traitement automatique, dont cependant le caractère systématique constitue un atout[1]. Or, le repérage sélectif des descripteurs, s'il est le fruit d'une lecture intelligente, répond surtout à une nécessité dans le cadre des pratiques manuelles où l'on ne peut envisager de lecture « intégrale » de tous les documents. Reste que, *a priori,* comme le fait remarquer Blair[2], il n'y a aucune limite, ni théorique ni pratique, au nombre de descripteurs : sur quels critères justifier alors la sélection de descripteurs ? L'intelligence de la lecture sélective ne devrait donc pas nécessairement se traduire par un repérage sélectif de descripteurs. Autrement dit, on voit mal pourquoi, dans la norme[3], l'une des phases de l'indexation est constituée par la « sélection des concepts », sinon pour une raison de faisabilité pratique : cet aspect de l'indexation ne relève donc pas à proprement parler du traitement documentaire mais de ses conditions de réalisation matérielle ; à ce titre, constitue-t-il encore un trait définitoire de l'indexation ?

Sur ce point, notre approche de l'indexation-extraction repose sur une prise en compte de l'intégralité du texte. Là encore, si un traitement automatisé facilite un repérage exhaustif des descripteurs, rien n'empêche de repérer, manuellement dans un texte, toutes les unités qui répondent aux critères du descripteur. Mais la nécessité s'exprime alors de façon aiguë de pouvoir, dans ce cas, s'en remettre à des critères de reconnaissance formels et pas uniquement à une intelligence de lecture.

- opposition 3 : indexation libre *versus* indexation contrôlée. Cette opposition repose sur deux critères : le premier concerne la présence ou pas d'un langage documentaire (dit aussi contrôlé). Le second, lié au premier, est plus implicite : il met en avant la confrontation ou pas à d'autres textes. Dans l'indexation contrôlée, le texte est analysé à travers le langage documentaire qui repose, lui, sur un ensemble de textes antérieurs ; dans l'indexation libre, l'examen du texte se fait sur la seule base du texte à indexer[4]. Dans les deux cas, il y a traduction mot à mot ; mais, dans un cas, la traduction est contrainte (contrôlée), dans l'autre, elle ne l'est pas : elle est libre. L'indexeur est libre d'utiliser les mots qu'il veut : il peut les extraire du texte comme choisir d'autres reformulations, mais c'est lui qui fait le choix des dénominations retenues.

  Pour cette raison, mais aussi du fait que l'indexation libre ne considère que le texte à indexer, l'indexation-extraction que nous proposons ne peut être assimilée à ce type d'indexation libre. L'indexation-extraction pourrait être

---

[1] Par exemple Rôle [1993, p. 137] présente l'indexation humaine comme « sélective » et, en cela « intelligente », qu'il oppose à l'indexation automatique, non intelligente, qui exige, elle, de « parcourir tout le document » mais qui présente l'avantage d'être « systématique ». Il semble ici que l'on confond le mode de lecture (sélectif pour l'indexeur humain) et le type de repérage des descripteurs (sélectif dans le sens où, parmi l'ensemble des descripteurs possibles, l'indexeur fait un choix).
[2] Blair 1990, p. 155.
[3] Norme Z 47-102 (1978).
[4] Neet 1990, p. 101 et suiv.

rapprochée de l'indexation « contrôlée » sur la base du renvoi qui se fait, *via* le langage documentaire, à d'autres textes, mais, comme nous l'avons vu, les textes dont sont issus les descripteurs ne sont pas ceux qui constituent la collection documentaire à indexer. En outre, l'opération de traduction sous-jacente à cette technique d'indexation se distingue radicalement de notre mode d'approche.

- opposition 4 : indexation post-coordonnée *versus* indexation pré-coordonnée. Cette opposition s'établit, elle aussi, sur la base de critères de nature différente. (i) Le premier concerne le langage documentaire utilisé : celui-ci peut être pré-coordonné (l'ordre des éléments est figé[1]) ou post-coordonné (l'ordre des éléments n'est pas contraint[2]).

  (ii) Le second concerne le type de termes d'indexation : sont dits pré-coordonnés ceux qui correspondent à des expressions (par exemple, l'expression « droit de la santé ») ; les termes d'indexation dits post-coordonnés correspondent à une séquence unique de caractères (par exemple, « droit », « santé »).

  (iii) Le troisième repose sur le lieu de la combinaison des descripteurs : si la combinaison des termes s'effectue au niveau de l'indexation, on parlera d'indexation pré-coordonnée, si elle se place au niveau de la recherche documentaire, on parlera plutôt d'indexation post-coordonnée.

  Le développement de l'informatique documentaire a, dit-on, favorisé l'indexation post-coordonnée puisqu'il devenait possible de proposer un nombre important d'unitermes combinables à la recherche par des opérateurs booléens[3] : la post-coordination relative à la combinatoire s'est trouvée confondue avec celle relative aux formes des descripteurs. Comme la norme en porte la marque, on a en effet assimilé le mode de combinaison des descripteurs (*a posteriori,* post-coordonné) avec la forme même des descripteurs (unitermes, termes dit post-coordonnés) : c'est ainsi qu'a été recommandée, dans la construction des thésaurus, l'adoption de « notion simple » devant correspondre à la forme de « mot simple[4] ». Or, on ne voit pas pourquoi on ne pourrait pas combiner par des opérateurs booléens aussi bien des unitermes que des expressions (« droit de la santé » et « sans domicile fixe », par exemple).

Cette typologie repose sur des critères de nature différente qui, tous appréhendés par la même notion de coordination (pré/post), ne laissent pas voir les différents phénomènes en jeu. En particulier, le type de coordination qui distingue les langages documentaires, s'il est proche de celui qui concerne les lieux de la combinatoire, n'entretient en revanche aucun rapport évident avec la coordination qui touche la forme des descripteurs. Le lien entre l'automatisation des systèmes d'information et l'adoption d'une morphologie spécifique des descripteurs (unitermes), mis en avant dans les discours normatifs, n'est pas, lui non plus, évident.

---

[1] C'est le cas du langage Rameau par exemple.
[2] C'est globalement le cas des langages documentaire de type thésaurus.
[3] Les opérateurs booléens sont en documentation de trois types : « et », « ou », « sauf ».
[4] Voir la norme Z 47-100 (1981).

Il paraît, sur ce point, nécessaire de reformuler les critères à la base de cette typologie. La coordination qui distingue les langages documentaires comme celle qui distingue les lieux de la combinaison des termes d'indexation relèvent plus du domaine de la recherche documentaire que de celui de l'indexation proprement dite. Reste la forme des descripteurs : qu'est-ce qui préside le choix de la forme des descripteurs (unitermes ou expressions) ? Quelle différence y a-t-il entre une indexation par unitermes (« droit », « santé ») et une indexation par expression (« droit de la santé ») ? Le point de vue normatif suggère que les deux types de descripteur sont équivalents et que les indexeurs peuvent opter pour l'un ou l'autre. Nous défendons ici, par la notion d'indexation-extraction, que la forme des descripteurs n'est pas indifférente et que ce ne sont pas des considérations pratiques, de nature informatique ou autre, qui peuvent déterminer la forme des descripteurs. Un point de vue linguistique permet de montrer les différences qui existent entre deux unitermes et l'expression dans laquelle ils sont éléments : les termes « santé » et « droit » ne sont pas équivalents au terme « droit de la santé ».

L'indexation que nous proposons privilégie les expressions au détriment des unitermes et pourrait à ce titre se comprendre comme une indexation de type « pré-coordonné » si l'on pouvait cependant « désintriquer » les différents aspects en jeu dans la notion de coordination.

En tâchant de situer notre approche dans le cadre des typologies classiques en indexation, nous avons pu relever que les critères d'opposition retenus n'étaient pas toujours explicites et présentaient en outre une certaine forme d'hétérogénéité. En guise de synthèse, on propose de mettre à plat les quatre typologies présentées ci-dessus et de se positionner par rapport aux critères implicites que nous avons dégagés :

Opposition	Critères	Indexation-assignation	Indexation-extraction
1	Agent humain	X	X
	Agent machinal	X	X
	Pratique professionnelle	X	X
	Champ d'application		X
2	Repérage sélectif	X	
	Repérage exhaustif		X
	Texte intégral	X	X
3	Langage documentaire	X	
	Confrontation textuelle	indirecte	directe
	Traduction	X	
4	Unitermes	X	X
	Expressions	X	X

En reformulant, pour des besoins de clarté, les typologies professionnelles de l'indexation, on remarque que notre approche de l'indexation conçue comme une extraction d'unités de discours s'écarte principalement de l'indexation classique sur les points suivants :
*(i) le mode de repérage est exhaustif (*versus *sélectif) ;*
(ii) pas de langage documentaire ;
(iii) pas de phase de « traduction » ;

*(iv) confrontation intertextuelle directe (*versus *indirecte).*

Les typologies classiques, même reformulées, laissent apparaître uniquement ces quatre types de différence. Les formes du descripteur en jeu ne semblent pas spécifiques à l'un ou l'autre type d'indexation ; même si les textes normatifs engagent à préférer les formes simples, les expressions constituent, de plus en plus, la forme dominante.

Toute la difficulté de notre approche du descripteur réside à montrer que, bien que les formes linguistiques que produit notre indexation-extraction, peuvent être très proches, voire les mêmes, que celles issues d'une indexation-assignation, la nature de ces formes et le rôle qu'elles sont destinées à jouer dans l'interprétation des textes sont radicalement différents de ceux que produit une indexation classique[1].

B - La recherche documentaire revisitée

Nous aborderons brièvement ce qu'implique, du point de vue de la recherche documentaire, notre approche du descripteur comme unité de discours.

Tout comme dans le modèle dominant de l'*Information Retrieval*, l'approche de la recherche documentaire que sous-tend l'indexation que nous proposons ne s'effectue pas sur la base d'une formulation spontanée[2] d'un « sujet de recherche[3] ». Là s'arrêtent les points de convergence : alors que la recherche documentaire classique se conduit à partir des mots du lexique, elle se conduit, dans notre cas, sur la base de mots de discours.

L'utilisateur se voit en effet proposer une liste de référents discursifs extraits des textes. Précisons à quoi peut ressembler une telle procédure, en reprenant les éléments que nous avons présentés dans le chapitre IV.

Nous avons établi la nécessité de distinguer deux modes de regroupement des sources selon le point de vue adopté (exploration et exposition). Autrement dit, la collection documentaire que nous proposons comprend deux types d'organisation des documents qui ne se superposent pas : une organisation par « domaines » qui constitue une forme d'exposition possible des documents et une organisation par « formations discursives » qui constitue une forme de regroupement possible des sources. Dans le cadre d'une telle collection documentaire, deux documents

---

[1] Dès le début de sa réflexion sur le descripteur, Le Guern [1984] a insisté sur la difficulté d'appréhension du descripteur comme unité de discours, son « apparence » pouvant le faire prendre pour un mot du lexique. La confusion est du même type que celle qui caractérise le terme de la terminologie : « On peut dire que le lexique concerne les mots indépendamment des choses, alors que dans la terminologie, les mots sont liés aux choses. Ils ont bien l'air d'être les mêmes, et beaucoup de gens s'y trompent, mais l'objet "mot" pertinent pour le lexique est une réalité totalement distincte de l'objet "mot" qui appartient à la terminologie », Le Guern 1989, p. 340.

[2] Notre conception de l'indexation ne suggère pas, contrairement aux tendances actuelles, que la recherche d'information doive idéalement passer par une formulation des besoins documentaires en « langage naturel », tout au contraire.

[3] Sans entrer dans le détail des problèmes que pose la notion de « sujet de recherche », nous mentionnerons simplement ici la distinction qu'établit Malrieu [1992] entre thème de recherche et thème de bibliographie, le second témoignant d'un effort effectué par l'usager pour adapter son thème de recherche à un système d'information. Dans ce paragraphe, c'est du thème de recherche que nous parlons.

exposés dans un même domaine n'appartiennent pas nécessairement à la même formation discursive.

L'utilisateur commence par sélectionner un domaine[1]. Cette sélection de domaine lui donne accès à une liste de termes de ce domaine ; la sélection d'un terme lui donne accès à un document (ou à une partie de ce document). La liste des unités de discours extraites de ce document est alors proposée : c'est dans ce réseau d'unités de discours que s'effectue la « navigation » dans un texte et le passage d'un texte à l'autre pour permettre la construction des thèmes de discours. Mais les textes rapprochés ici par les unités de discours sont ceux qui appartiennent à une même formation discursive et non plus au même domaine. C'est sur ce type de configuration qu'il nous semble envisageable de penser la recherche documentaire. Cette esquisse trouve des échos dans deux types de travaux portant sur les systèmes de recherche d'information.

La recherche documentaire sous forme de « navigation » entre unités de discours est en effet à l'œuvre dans les prototypes réalisés par les membres de l'équipe SYDO[2]. Nous y reviendrons dans le dernier paragraphe de ce chapitre. Par ailleurs, l'approche de la recherche documentaire esquissée ici sur la base d'une conception du descripteur comme unité de discours rejoint les conclusions de travaux réalisés, dans une perspective cognitive, sur les situations de recherche documentaire. Ainsi, Kolmayer[3] par exemple montre-t-elle que la situation de recherche documentaire ne saurait être réduite à une indexation de requêtes et qu'il est, en ce sens, nécessaire de revoir non seulement l'ergonomie classique des systèmes de recherche documentaire mais aussi les types de descripteurs issus du processus d'indexation.

Cette rapide esquisse d'un nouveau type de recherche documentaire souligne qu'il n'est pas aussi nécessaire que semblaient l'indiquer les normes de connaître les « intérêts » des utilisateurs, ou encore la terminologie qu'ils emploient, pour pratiquer l'indexation. La recherche documentaire peut se faire par le biais d'un guidage progressif à l'intérieur d'une collection documentaire, guidage qui emprunte aux textes eux-mêmes leurs éléments.

Une telle approche de la recherche documentaire suppose un accès au texte intégral dès le stade de la recherche elle-même, et non plus au seul niveau de la consultation de textes. Là encore une nécessité « théorique », liée au fonctionnement linguistique du processus interprétatif, trouve une voie de réalisation pratique dans les « nouvelles technologies », qui n'était guère envisageable si nettement aux premiers temps de l'indexation ou de la recherche documentaires. Mais, *a contrario*, on réalise à quel point, dans les descriptions normatives de la recherche documentaire, les facteurs de faisabilité matérielle prédominent, parfois au détriment d'autres aspects (linguistiques ou cognitifs).

---

[1] Un texte peut appartenir à plusieurs domaines à la fois ; l'appartenance d'un texte à un domaine se fait sur la base d'un dépouillement terminologique. Cet aspect du dépouillement terminologique pose plusieurs types de problèmes que nous nous contentons de lister ; on renvoie aux travaux dans ce domaine qui permettent d'envisager la réalisation d'une telle classification de façon automatique [par exemple Perron 1988 et 1991]. Les principaux problèmes du dépouillement terminologique sont les suivants : comment reconnaître les termes dans un texte ? Comment décider de l'appartenance d'un terme à un domaine ? Quel lien y a t-il entre l'appartenance d'un terme à un domaine et l'appartenance d'un texte (ou d'un segment de texte) d'où est extrait ce terme à un domaine ?

[2] Sur ce point, voir par exemple Kuramoto [1995 et thèse en cours].

[3] Kolmayer 1997.

La recherche documentaire que nous proposons s'effectue sur la base d'un chaînage entre unités de discours : toutes les unités linguistiques ne constituent pas au même titre de bons maillons pour ce que nous appellerons les « chaînes de référence ». La problématique du descripteur comme unité de discours si elle suppose que tout descripteur est nécessairement une unité de discours ne valide pas pour autant la réciproque. C'est cette restriction que nous examinerons ci-après.

Dans ce paragraphe, nous nous sommes attachée à déterminer les conséquences de l'approche du descripteur comme unité de discours sur les processus d'indexation[1] *et de recherche documentaires, en essayant de situer nos propositions par rapport aux pratiques les plus courantes. Certes,* a priori, *un écart important se marque entre notre conception de l'indexation et les pratiques courantes, mais il reste délicat de circonscrire précisément cette différence : elle apparaît plus nettement dans le cadre du modèle du thème discursif (indexation explicative versus indexation interprétative) que dans le cadre des typologies classiques : sur ce point encore, on relève que le seul référentiel documentaire reste insuffisant pour manipuler, en toute généralité, les faits d'indexation.*

Cependant l'étude des typologies existantes a été nécessaire pour souligner le rôle des facteurs techniques dans les approches de l'indexation. Sur ce point, on a pu remarquer que nos propositions trouvaient un moyen d'application optimal en recourant à certaines facilités offertes par la technologie actuelle : l'exploitation de la capacité de stockage de l'ordinateur rend moins nécessaire la sélection des descripteurs ; la possibilité de présenter, dès le stade de la recherche documentaire, les textes intégraux des sources rend envisageable l'approche de la recherche documentaire en termes de parcours interprétatif ; la possibilité de recourir à des analyseurs linguistiques permet de concevoir l'indexation comme une extraction automatisée de référents discursifs. Mais, comme nous avons cherché à le montrer tout au long de cette recherche et comme tentera encore de le souligner notre étude du descripteur, l'exploitation de la technologie est ici mise au service d'un projet d'indexation établi sur des fondements théoriques.

En creux et a contrario *se dessine, dans les approches classiques de l'indexation et de la recherche documentaires, le poids des contraintes matérielles qui tend à obscurcir une approche du processus de l'indexation lui-même. Les définitions normatives apparaissent sur ce point comme la rencontre de multiples facteurs de nature hétérogène. On comprend mieux non seulement certains traits posés comme définitoires de l'indexation (comme la sélection de descripteurs par exemple) mais aussi le vocabulaire employé par la norme. À ce titre, la notion de « représentation » est à comprendre aussi*[2]*, semble-t-il, dans un cadre où les textes indexés ne sont pas immédiatement accessibles. Ainsi se mêlent, à des visions particulières du lexique et de la référence, des aspects beaucoup plus pragmatiques, chacun de ces éléments hétérogènes formant un tout, ayant tendance à constituer un ensemble* a priori *cohérent mais fortement opaque.*

---

[1] Réduite au seul aspect de la détermination des descripteurs.
[2] Voir notre interprétation de la notion de « représentation » proposée dans le chapitre II de cette recherche (§ I.1).

I.2 - Restriction : toute unité extraite du discours n'est pas nécessairement un descripteur

On pourra trouver que notre approche du descripteur comme unité extraite du discours réalise un retour en arrière dans les années 50, au moment des premières tentatives d'indexation automatique fondée sur l'extraction des unités mêmes du texte.

Si le principe d'extraction est le même que celui proposé par Luhn ou Taube, le type d'unités qu'il s'agit d'extraire est cependant radicalement différent. Sur ce point, il importe d'examiner les arguments avancés contre l'indexation-extraction. On pourra y identifier des confusions, souvent faites, entre le processus d'extraction lui-même et le type d'unités concerné (I.2.1). En outre, l'examen de ces arguments nous permettra d'établir, dans ces grandes lignes, le cahier des charges du descripteur comme unité de discours (I.2.2).

I.2.1 - EXAMEN DES ARGUMENTS CONTRE L'INDEXATION-EXTRACTION

Sans nous livrer ici à une étude exhaustive des arguments habituellement présentés par les professionnels de l'information contre l'indexation-extraction, nous proposerons un aperçu des principales critiques. Nous montrons d'abord (A) que les arguments avancés se situent généralement en « porte-à-faux » du problème, notamment parce que ne sont pas prises en compte, dans l'évaluation de l'extraction, les particularités de la procédure (elle exploite les textes eux-mêmes). Nous évoquons, ensuite (B), les trois principaux problèmes qui, aux yeux des praticiens, empêchent de considérer l'extraction comme un bon principe pour l'indexation ; là encore, parmi les arguments avancés, se marque la présence d'un état, aujourd'hui dépassé, de la technologie.

A - Des arguments en porte-à-faux

Le plus souvent, comme le rappelle Chaumier[1], le terme « indexation » est opposé à celui d'« extraction ». Le résultat lui-même de l'une ou l'autre opération est nommé différemment : si l'indexation produit des descripteurs (des termes *désignés* par l'indexeur, dans la terminologie de Neet[2]), l'extraction, propose, elle, des « mots-clés[3] » (des termes *dérivés* du texte chez Neet). Cependant, malgré les différences terminologiques établies, les fonctions attendues de l'un ou l'autre type d'unité d'indexation sont rigoureusement identiques. En effet, si le mot-clé présente la particularité d'être extrait d'un texte, il doit, au même titre que le descripteur, « caractériser » le contenu du document et permettre de le retrouver. Autrement dit, comme l'a particulièrement mis en valeur Michel Le Guern[4], le mot extrait du discours est assimilé au mot issu d'un lexique documentaire ; ni le processus de l'indexation ni celui de la recherche documentaire ne sont repensés en fonction de cette différence[5].

---

[1] Chaumier 1989, p. 15.
[2] Neet 1990, p. 102.
[3] Le mot-clé est un « mot choisi dans le titre ou le texte d'un document, caractérisant son contenu et permettant la recherche de ce document », norme Z 47-102 (1978), p. 231.
[4] Le Guern 1984.
[5] Sans doute à cause de la difficulté qu'il y a de percevoir, sans le recours à un prisme théorique, la différence de statut entre ces unités.

Or nous avons essayé de montrer précédemment, d'une part, que le statut du descripteur comme mot extrait du discours impliquait une autre appréhension et de l'indexation et de la recherche documentaires, et d'autre part, que selon le type d'indexation visé (explicatif *versus* interprétatif), la morphologie du descripteur était différente.

L'approche de l'indexation-extraction par la notion de mot-clé souligne encore, sous une autre forme, ce que nous avons pu noter à plusieurs reprises dans cette étude : l'indexation se pense exclusivement dans sa dimension lexicale, au niveau des mots eux-mêmes, des descripteurs, sans prise en compte des discours auxquels ils sont liés ou dont ils sont issus. Cette absence de prise en compte de la spécificité du mot-clé comme extrait du discours conduit à des évaluations erronées. On s'étonne en effet que le mot extrait du discours ne fonctionne pas comme le mot extrait d'un lexique documentaire et on condamne alors la procédure de l'extraction elle-même, sans avoir pris la peine de mesurer ce que supposait de disposer de descripteurs extraits de discours. Ainsi cette fin de non-recevoir formulée à l'égard du système Stairs[1] : « Les recherches faites en utilisant le logiciel employé (STAIRS) n'ont permis de retrouver que 20% des documents pertinents. Il est donc évident que la plus grande partie des documents pertinents ne contenaient pas les mots et les phrases utilisés dans les questions posées, en dépit du fait que les documents pertinents non retrouvés concernaient des sujets qui intéressaient les chercheurs. Comme les techniques d'indexation automatique sont presque toujours fondées sur l'extraction de vocabulaires, il n'est nullement évident que ces techniques, si complexes soient-elles, puissent être efficaces en vue de fournir des représentations adéquates des documents.[2] »

Il apparaît clairement, dans cette citation, que la modification introduite dans le mode de détermination des descripteurs ne s'est en rien répercutée ni sur la perception de l'indexation (il s'agit toujours de représenter le contenu) ni sur celle de la recherche documentaire (on retrouve le modèle de l'*Information Retrieval*) : le modèle de référence implicite reste toujours celui édicté par la norme.

B - Des confusions entre procédure d'extraction et type d'unité extrait

De façon générale, les professionnels défendent la nécessité d'une indexation « contrôlée » en mettant en avant les problèmes rencontrés par la seule extraction de mots de discours.

Les problèmes le plus souvent cités sont les suivants[3] : l'indexation-extraction présenterait le désavantage (i) de ne gérer ni la synonymie, ni l'homonymie ; (ii) elle ne pourrait identifier les thèmes « cachés » ; (iii) elle ne détecterait pas non plus ni les termes composés ni la diversité flexionnelle des formes.

Nous avons regroupé les différents types de problèmes classiquement évoqués en trois ensembles qui nous semblent mettre en jeu des aspects différents, ici tous appréhendés sous le seul angle de l'extraction.

On peut remarquer que le groupe (i) de problèmes ne concernent pas directement et uniquement l'indexation. Elles concernent aussi et surtout la recherche

---

[1] Système créé dans les années 70 par IBM.
[2] Blair 1986, vol. 13, p. 23 cité *in* de Grolier 1988, p. 472.
[3] On reprend ici ceux synthétisés dans Neet 1990, p. 106 et suiv.

documentaire vue sous l'angle du seul modèle de l'*Information Retrieval* . En effet, dans l'approche que nous proposons, les « synonymes » constituent, dans un texte, autant de relais lexicaux qui permettent la construction d'un thème de discours. De même, l'homonymie, qui pose surtout des problèmes d'interprétation quand les mots sont appréhendés hors contexte (par exemple, le mot « avocat »), ne constitue pas un obstacle dans une approche qui privilégie le discours.

Le problème identifié en (ii) renvoie à la notion de thème inféré mise en valeur par Marandin[1]. Cette approche montre que si un thème n'est pas toujours « nommé » dans un texte, il y a toujours des indices, des ancrages discursifs qui permettent de l'inférer[2]. Là encore, ce sont ces indices-là que l'indexation peut se proposer d'extraire. Le problème de l'absence d'extraction de « thème inféré » ne se pose que dans le cadre d'une indexation de type interprétatif qui cherche à déterminer des thèmes ; or, on a vu qu'elle n'était pas la seule envisageable ; nous pensons en outre qu'elle n'est pas non plus la plus pertinente.

Les deux problèmes regroupés en (iii) posent, eux, crucialement, le problème du type d'unités extrait. Si les unités extraites sont des unités linguistiques, c'est-à-dire spécifiées d'un point de vue linguistique, l'extraction alors réalisée à partir d'un modèle linguistique ne produira aucun de ces deux problèmes[3]. Ces problèmes se posent quand l'extraction se réalise en dehors de tout modèle formel du descripteur. Sur ce point, on peut remarquer que l'essentiel des arguments avancés contre l'indexation-extraction repose sur un modèle de l'extraction qui date des années 50 : le modèle de l'indexation par extraction d'unitermes proposé par Taube en 1952. Ce modèle, qui ne met en jeu aucun savoir de nature linguistique, ne peut appréhender le texte que comme un ensemble de chaînes de caractères. Les unités qui sont extraites n'ont pas nécessairement un statut linguistique de « mot de discours ».

L'extraction par unitermes revient à proposer, par exemple, comme descripteurs, les suites « droit » et « travail », alors que le texte exhibe, lui, une unité de discours de la forme « le droit du travail ». De même, une telle méthode d'extraction par chaîne de caractères ne pourra rapprocher les formes flexionnelles d'une même unité linguistique (les suites « cheval » et « chevaux » resteront irrémédiablement séparées). Autrement dit, l'ensemble des problèmes regroupés en (iii) interroge non pas la procédure d'extraction elle-même mais les types d'unité que l'on cherche à extraire ; c'est cependant la procédure elle-même qui est remise en cause, comme le montre par exemple cette condamnation du « full text » : « The ambiguity of uncontrolled language expressions is not resolved in full text system. This is a source of noise in retrieval in those case where disambiguisation is possible. For example, "silver" will be found in "silver jubilee", "silver fir", "German silver", etc. Noise in full text information retrieval system can be excessively large.[4] »

L'indexation-extraction est ici, et comme dans la majeure partie des cas, envisagée sous la seule forme d'une extraction d'unitermes, ou plutôt de chaînes de caractères. Or, comme nous le préciserons ci-après, toutes les chaînes de caractères d'un texte ne constituent pas des unités de discours susceptibles d'être des descripteurs. Là encore se marque la nécessité, aussi bien pour évaluer les procédures que pour automatiser un traitement documentaire, de se doter d'un

---

[1] Voir le chapitre II § II.1.1.
[2] Voir Marandin 1988 pour plus de précision.
[3] Voir, sur ce point, le paragraphe III.3 de ce chapitre.
[4] Fugmann 1993, p. 99.

modèle explicite du descripteur, faute de quoi, la critique risque de manquer de portée et l'indexation-extraction de pertinence.

C'est ainsi qu'à notre sens, les arguments regroupés en (iii), s'ils sont trop massivement reportés contre l'indexation-extraction, n'en soulignent pas moins la difficulté de l'extraction d'unités de discours : elle impose nécessairement un traitement des flexions et des procédés de repérage des « termes composés ». Nous y reviendrons ci-après.

Notons que les praticiens de la documentation et des bibliothèques ne sont pas les seuls à se montrer déçus par les systèmes d'indexation automatisés et en cela souvent opposés aux techniques d'extraction textuelle. Tous les professionnels qui se trouvent concernés, pour une raison ou pour une autre, par le « filtrage de l'information » se heurtent au même problème lié à l'extraction de chaînes de caractères. Ainsi, dans le contexte actuel de l'utilisation d'Internet, certains États, souhaitant interdire l'accès sur leur territoire à des sites véhiculant des propos usuellement condamnés par leurs lois nationales (pour la France, propos négationnistes ou pédophiliques par exemple), se tournent vers l'utilisation de « logiciels de filtrage » qui, sur la base d'une liste de mots-clés, pourraient bloquer automatiquement l'accès à des sites Internet. La solution, testée, a dû être abandonnée. Pour reprendre l'exemple cité dans *Le Monde*[1], les logiciels de filtrage testés en venaient à interdire l'accès au site de la Maison Blanche sur la base de la chaîne de caractères « couple » apparaissant dans la suite « le couple présidentiel ». D'autres exemples, comme les aberrations produites par la détection de la suite « sein » (qui se trouve dans la locution « au sein de »), ont conduit à privilégier des systèmes d'étiquetage manuels des sites. Mais là encore, comme le soulève l'article du *Monde*, dans le « vocabulaire standard » utilisé, on ne précise pas « le contenu des étiquettes », c'est-à-dire les objets visés par les mots retenus. On le voit : la procédure d'extraction, pour être appliquée avec quelque succès, nécessite pour le moins un examen attentif des unités de discours elles-mêmes.

On pourra noter enfin que les systèmes récents qui proposent une extraction de « termes composés », ou encore de groupes ou de syntagmes nominaux, à des fins d'indexation (mais aussi pour réaliser d'autres objectifs professionnels), ne sont pas toujours de nature à pouvoir restaurer la confiance des professionnels. En effet, comme a pu le montrer Sophie David[2], nombre d'entre eux effectuent une extraction sans se doter préalablement d'un modèle explicite de l'unité qu'ils cherchent à identifier dans un texte. Les critères d'extraction restent alors le plus souvent implicites ; ceux qui sont exprimés apparaissent très *ad hoc*. Les procédures d'extraction ne faisant pas toujours appel à un savoir de nature linguistique, le statut des unités obtenues est pour le moins incertain : on n'est pas toujours sûr de disposer d'unités de discours exploitables en contexte professionnel.

Cet examen rapide, non exhaustif, des principaux arguments avancés par les professionnels contre l'indexation-extraction souligne les difficultés d'une telle entreprise :

- *d'une part, l'indexation-extraction oblige à sortir du cadre d'appréhension classique de l'indexation et de la recherche documentaires ; l'extraction*

[1] *Le Monde* du 4/07/1997, p. 21 : « La technologie devient l'unique recours des Américains pour filtrer le contenu d'Internet ».
[2] David 1993a, chapitre V, p. 192-251.

d'unités de discours en soi ne vaut rien si elle ne s'accompagne pas d'une réflexion plus globale sur les particularités de l'interprétation des discours ;

- *d'autre part, l'indexation-extraction oblige à établir un modèle formel du descripteur ; cette modélisation des unités de discours est loin d'être, en linguistique comme en indexation, triviale.*

Ainsi avons-nous cherché à faire apparaître les particularités de l'indexation-extraction que nous proposons :
(i) elle appelle un modèle de représentation linguistique du descripteur ;
(ii) elle résiste aux arguments habituellement présentés par les professionnels contre l'extraction en indexation.

I.2.2 - « CAHIER DES CHARGES » DU DESCRIPTEUR COMME UNITÉ DU DISCOURS

Si, comme nous l'avons montré, le descripteur est nécessairement une unité extraite du discours, toute suite extraite du discours ne peut être considérée comme un descripteur. En effet, la construction de la référence ne s'établit pas, comme nous l'avons vu dans la première partie de cette recherche, sur la base d'un « mot » seul (encore moins sur la base d'une suite de caractères qui n'aurait pas même le statut de mot) ; par ailleurs, elle n'est pas le fruit non plus d'une relation directe entre les mots et les choses.

Notre étude de la problématique du descripteur en indexation reprendra, sur ce point, l'ensemble des remarques que nous avons faites précédemment sur le rapport entre sens et référence et sur la construction de la référence en discours.

La problématique du descripteur que nous développerons ici s'appuie sur l'approche discursive de l'indexation que nous avons présentée au chapitre IV. Il faut en effet considérer que :

- les textes sont regroupés dans une collection documentaire sur la base d'une formation discursive ou, au moins, sur la base d'un *a priori* qui rend valide une analyse de type interdiscursif ;

- dans ce cadre, le descripteur fonctionne comme « terme-pivot » dans un texte et comme « terme-relais » entre textes : ces deux dénominations signalent que le descripteur se conçoit comme un référent discursif à même de construire un thème de discours[1].

On sait qu'un thème de discours, une fois construit et nommé, produit un effet de stabilité référentielle (ou encore de continuité thématique) : l'enjeu du descripteur consiste à permettre de créer cette stabilité référentielle. Pour étudier la morphologie du descripteur sous cet angle, il importe de considérer les caractéristiques des textes eux-mêmes. En effet, comme l'a par exemple clairement mis en valeur Corblin, la stabilité référentielle créée en discours ne correspond jamais à une stabilité linguistique, ou encore, la référence au « même objet » ne s'effectue jamais, dans un texte, par les mêmes mots : « Rien n'est plus opposé au fonctionnement des langues naturelles que l'expression de l'identité référentielle par

[1] Les notions de « terme-relais » et de « terme-pivot » sont précisées dans les paragraphes II et III de ce chapitre.

la répétition littérale d'un identificateur. [...] La forme typique, presque canonique, de la chaîne de référence naturelle est au contraire une relation entre termes dissemblables.[1] »

Pour capter, plus précisément, les éléments qui, de nature forcément hétérogène, permettent à un utilisateur de créer un objet de discours, nous recourrons à la notion de « chaîne de référence ». L'étude, menée à la fois d'un point de vue logique et linguistique, des éléments qui constituent cette « chaîne » nous permettra de déterminer la morphologie tant d'un point de vue logique que linguistique du descripteur (§ II et III de ce chapitre).

I.3 - Enjeu du descripteur : la construction de chaînes de référence

Nous présenterons tout d'abord la notion de chaîne de référence telle que les linguistes ont pu l'utiliser (I.3.1) ; nous discuterons ensuite la pertinence de cette notion pour notre définition du descripteur (I.3.2). Nous spécifierons enfin les types d'unité linguistique régulièrement mis en cause dans la construction de chaînes de référence (I.3.3), à partir de quoi nous pourrons engager notre étude sur la morphologie du descripteur.

I.3.1 - PRÉSENTATION DE LA NOTION DE CHAÎNE DE RÉFÉRENCE

Nous rappelons d'abord la notion telle qu'elle a été formulée d'un point de vue logique ; nous privilégions ensuite la lecture linguistique qui a pu en être faite.

A - Approche logique des chaînes de référence

La notion de chaîne de référence, que l'on peut, en première approximation, définir comme « la suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence[2] » a été établie par le philosophe (ou plus précisément le logicien) Chastain[3] dans le cadre d'une réflexion plus globale sur la « référence singulière[4] ».

De façon schématique, on peut dire que son apport consiste à approcher la référence des termes singuliers en termes de construction d'objet dans un espace contextuel[5] : c'est au terme d'un « parcours » dans un contexte donné qu'un terme singulier établit une relation référentielle. Il importe donc de circonscrire, dans un texte, ces parcours, ces tracés, en encore les contextes, dans lesquels la référence peut être considérée, d'un point de vue interprétatif, comme stabilisée, construite.

---

[1] Corblin 1995, p. 174.
[2] *Ibid.*, p. 27.
[3] Chastain 1975.
[4] La référence de ce que les logiciens nomment, depuis Quine, les termes singuliers (généralement, les noms propres, les descriptions définies, les pronoms personnels, les démonstratifs). Chastain [1975, p. 198] définit ainsi les termes singuliers : « Singular terms are the ones whose role is to be referentially connected with objects ».
[5] Chastain s'oppose aux théories logiciennes « classiques » de la référence, comme celle de Russel, et reprend une critique qu'avaient entamée Strawson et Donnellan : « Previous theories have generally tried to explain the connection between a singular term and its referent as a function of the meaning of the term and the properties of the referent, paying little or no attention to the circumstances in which the term is uttered », Chastain 1975, p. 196.

C'est par la notion de chaîne de référence que de telles zones peuvent être identifiées.

Dans l'approche de Chastain, outre que plusieurs chaînes de référence peuvent se trouver dans un même texte[1], il est nécessaire de distinguer deux types de chaîne de référence selon le type de lien contextuel mis en cause. Les chaînes anaphoriques s'établissent dans un même contexte tandis que les chaînes référentielles s'établissent, elles, entre différents contextes[2] : « We have *anaphoric chains within contexts*, such as if one expression in the chain refers to a given thing then so do all the others, and we have also *referential chains between contexts*, for which the same conditions hold : if one expression in the chain refers to a given thing then do so all the others.[3] »

La référence se construit en discours sur la base de l'entrecroisement de plusieurs chaînes appartenant à ces deux types : « Since the links in referential chains connecting different contexts are also links in anaphoric chains within contexts, we can see how very lenghtly referential chains can be constructed : a singular term T_1 in a context C_1 is anaphorically linked with another singular term T_2 in C_1, which is in turn linked with T_3 in C_1, which is referentially linked with T_4 in another context C_2, which is anaphorically linked with T_5 in C_2, which is referentially linked to T_6 in C_3, and so on, *until we come at last to some singular terme T_n which refers to some object, which thus counts also as the referent of all singular terms along the whole chain from T_1 to T_n.*[4] »

Telle que Chastain la décrit, la notion de chaîne de référence fait apparaître que la construction d'un objet de discours (le T_n dans la citation) – l'objet qui permet, lui, un découpage du référent mondain –, s'effectue par le biais de saisies multiples de référents discursifs (T_1-T_6), en suivant un processus complexe mais repérable, complexe dans la mesure où la notion de chaîne passe outre les frontières de la phrase[5], repérable par le type de contexte et d'unités linguistiques mis en jeu.

Par la notion de chaîne de référence, Chastain établit une distinction entre les concepts logico-sémantiques de « dénotation » et de « référence[6] », ou plutôt il montre que dénotation et référence ne sont pas nécessairement coïncidentes : « Unique denotation doesn't determine reference in the case of a definite description in an overt discourse which forms a referential chain with a singular term in an antecedent covert discourse. The reason for this is that the connection with the term in the other discourse provides *a way* for the description to be connected with a referent *independent of whether its descriptive content happens to*

---

[1] Chastain 1975, p. 267.
[2] Il y a aussi, dans les textes, des types de contexte qui ne participent pas à la construction de la référence ; Chastain les nomme « contextes isolés » : « If a singular term in a given context is not referentially linked to a singular term in another context, I will call it *referentially isolated* », *Ibid.*, p. 214.
[3] *Id.* (C'est nous qui soulignons).
[4] *Id.* (C'est nous qui soulignons).
[5] *Ibid.*, p. 216.
[6] Nous ne détaillerons pas ici, d'un point de vue logique, la différence que suggère Chastain entre dénotation et référence. Nous en proposons une lecture linguistique ci-après, où « dénotation » renvoie à la notion de « référence virtuelle » et « référence » à celle de référence discursive.

fit that referent. Denotation can fail and reference succeed because *there is an alternative route to the thing referred to.*[1] »

S'il y a des cas où la référence se distingue de la dénotation (à chaque fois qu'il y a « chaîne[2] »), il y a aussi des cas où il n'y a pas de référence (pas de « chaîne ») : seule la dénotation est alors à l'œuvre dans la construction référentielle.

La nécessité d'une notion comme celle de chaîne de référence apparaît donc clairement : connaître la « dénotation » d'un terme ne suffit pas toujours à comprendre sa référence en discours.

L'essentiel de la théorie de Chastain repose sur la notion de contexte ; or celle-ci est, pour toute lecture qui n'est pas logicienne, trop lâche. En effet, Chastain définit un contexte comme « anything that has meaning or sense [...]. Anything which expresses something or represents something is a context[3] *». Comme le souligne Corblin*[4]*, au vu des exemples donnés par Chastain, on pourrait assimiler la notion de contexte à celle de discours. À ce titre, on dira qu'une relation anaphorique s'établit à l'intérieur d'un discours et que les chaînes référentielles s'établissent entre discours. Cependant, il y a des cas, notamment des cas d'utilisation du nom propre dans un discours, où il faut postuler l'existence d'une chaîne référentielle à l'intérieur d'un même discours*[5]*. La notion de contexte n'est donc pas toujours suffisante pour identifier la présence d'une chaîne et le type de chaîne en cause. C'est notamment par la proposition d'autres critères que se signale l'apport des linguistes.*

B - Approche linguistique des chaînes de référence

Les linguistes se sont intéressés à la notion de chaîne de référence pour l'étude de la construction référentielle en discours[6] et, plus précisément, pour repenser la notion linguistique d'anaphore[7].

La notion de chaîne de référence permet de dépasser la frontière de la phrase et de penser l'anaphore en termes de dépendances interprétatives[1]. La nécessité de

---

[1] Chastain 1975, p. 237-238 (C'est nous qui soulignons).
[2] « When a term is referentially linked with a previous context, denotation is neither necessary nor sufficient for reference », *Ibid.*, p. 238.
[3] *Ibid.*, p. 195. De ce point de vue, un discours par exemple constitue un contexte. La difficulté que pose, nous semble-t-il, la théorie de Chastain, c'est que, si elle prend pour exemple privilégié la référence en discours, le propos se veut en réalité plus général (la référence singulière en général) : tous les éléments de sa théorie ne sont donc pas définis sur le même plan ; ainsi de la notion de contexte qui est à entendre dans un sens large (Chastain 1975, p. 195 : « A discourse is a context [...] a map is a context [...] a picture is a context [...] my visual field is a context »). Sur ce point, voir Corblin 1995, p. 28 : « Chastain [1975] cherche à déduire des relations qui s'établissent au sein du discours un modèle pour concevoir la référence singulière ».
[4] Corblin 1995, p. 154-156.
[5] Corblin [1995] qui étudie les chaînes de référence dans les romans constate en effet que l'« acte de baptême » qui constitue la référence d'un nom propre ne fait pas appel, dans le roman, à d'autres discours. Bien que le cas du roman ne soit pas central en indexation, nous prenons en compte cette remarque de Corblin.
[6] Par exemple Corblin 1987, p. 8 : « Au lieu de considérer une relation entre un univers d'objets et des expressions capables de les désigner, on se représente plutôt le phénomène comme une mise en relation de mentions dans une séquence ».
[7] Notamment Corblin [1987, 1995, par exemple] ou Marandin [1988, 1997, par exemple].

repenser la notion classique d'anaphore repose sur le constat que, dans les langues naturelles, et contrairement aux langages formels, l'identité référentielle perçue d'un point de vue interprétatif se fonde sur un système de reprises qui emprunte des éléments hétérogènes, à la fois d'un point de vue formel (type d'unité) et d'un point de vue référentiel (type de dénotation) : « Dans les textes en langue naturelle, la mention d'entités ou d'objets se réalise par la construction de chaînes d'identité ou d'association entre des *segments formellement et interprétativement hétérogènes*. C'est là une caractéristique des langues naturelles qui s'opposent aux traitements automatiques immédiats. Il est facile dans un texte de repérer des segments formellement identiques (strictement ou modulo une règle d'équivalence fixe) mais les chaînes de référence, de manière typique et pourrait-on dire constitutive, ne reposent pas sur l'identité formelle. Rien n'est plus opposé au fonctionnement des langues naturelles que l'expression de l'identité référentielle par la répétition littérale d'un identifieur absolu comparable à l'usage de symboles de constante dans les énoncés mathématiques.[2] »

La notion de chaîne de référence souligne le paradoxe suivant : la permanence référentielle ne correspond jamais, dans les langues naturelles, à une permanence linguistique. De là peut se dégager une loi, que Corblin propose d'appeler « loi de diversité formelle » de l'identité référentielle en discours. Cette spécificité de la construction du référentiel textuel est liée à la nature même de l'objet textuel, à ce que Corblin nomme « la plasticité des objets textuels » et qui rend compte du fait que « dans un texte, la permanence référentielle s'accomplit au moyen de saisies et de ressaisies des objets qui sont constitutives de l'interprétation[3] ». Autrement dit, il peut y avoir chaîne, c'est-à-dire interprétation d'une identité référentielle, entre des éléments qui n'ont pas la même valeur référentielle[4] : ce n'est pas exactement le même objet qui est construit en divers points d'une chaîne de référence.

L'application de la notion de chaîne de référence dans le contexte de l'analyse linguistique suppose une reformulation des notions de chaîne anaphorique et chaîne référentielle, dans la mesure où la notion de contexte, dominante dans le modèle de Chastain, reste insuffisante pour une approche linguistique.

Corblin propose une révision en ces termes :

- les chaînes anaphoriques établissent des connexions référentielles sur des *bases linguistiques* : le calcul interprétatif qui permet d'établir l'identité référentielle est « déclenchée et régie par le contenu linguistique de la forme[5] » ;

---

[1] Ce qui permet d'élargir la notion classique d'anaphore. Deux groupes nominaux (ou plus) pouvant être compris comme « anaphoriques » l'un de l'autre, comme dans l'exemple suivant [Kafka, *Le Procès*, cité *in* David 1989, p. 100] : « On racontait à ce sujet une anecdote qui paraissait fort vraisemblable : *un vieux fonctionnaire*, paisible et brave homme s'il en fut, avait étudié sans répit pendant un jour et une nuit – car *ces employés* sont extrêmement laborieux – une cause des plus épineuses ». La relation d'anaphore n'est plus crucialement distinguée de celle de co-référence, comprise comme « une relation entre deux expressions qui réfèrent au même particulier sans être connectés par une relation *linguistique* d'anaphore », Corblin 1995, p. 166 (c'est nous qui soulignons).
[2] *Ibid.*, p. 174 (c'est nous qui soulignons).
[3] *Ibid.*, p. 191.
[4] David 1989, p. 100.
[5] Corblin 1995, p. 168.

- les chaînes référentielles établissent des connexions référentielles sur des *bases communicatives* : « des expressions sont associées à un designatum en vertu de connaissances contingentes, et si nous les traitons comme équi-référentes, c'est uniquement en fonction de connaissances constituées dans la communication[1] ».

On peut illustrer la différence entre types de chaînes par les deux exemples suivants :
(1) *Nixon* réfléchissait. *Il* était dans une situation périlleuse.
(2) *Nixon* réfléchissait. *Le président des USA* était dans une situation périlleuse.

En (1), la connexion référentielle est linguistiquement régie (relation classiquement anaphorique antécédent/pronom) : (1) constitue une chaîne anaphorique. En (2), le lien entre « Nixon » et « le président des USA » s'établit en vertu de notre connaissance de l'univers de référence : (2) constitue une chaîne référentielle. On peut en effet utiliser « indépendamment » les deux expressions « Nixon » et « le président des USA », sans savoir ce qui, historiquement, à un moment donné, les a reliés, alors qu'en (1), les deux expressions ne sont pas de la même façon interprétativement indépendantes.

Comme le note Corblin, les chaînes référentielles, établies sur des bases communicatives, ne sont pas sans rappeler les « chaînes communicatives » qui caractérisent, dans le modèle des mondes possibles de Kripke, le fonctionnement référentiel du nom propre[2].

Si Corblin propose de distinguer chaînes anaphoriques et chaînes référentielles sur la base du type de calcul qui permet de dériver l'identité référentielle (calcul de type linguistique *versus* calcul de type communicatif), il explicite également ce qui les rapproche : le type de lien actif dans les chaînes.

Dans chacun des deux types de chaîne, l'identité référentielle peut s'établir entre éléments de deux façons :

- par la mention des mêmes entités : dans une chaîne anaphorique, ce cas peut être représenté par une succession de pronoms (La jeune fille... *Elle... Elle...*) ; dans une chaîne référentielle, par une succession de noms propres (Le président des USA... *Nixon... Nixon...*).

- par la mention d'entités différentes en situation de dépendance interprétative : dans une chaîne anaphorique, on pourra avoir par exemple des suites comme « La maison... la porte d'entrée... la fenêtre » ; dans les chaînes référentielles, on aura par exemple la suite « Flaubert... l'ermite de Croisset ».

Corblin en vient à établir un schéma des connexions référentielles en discours qui se donne sous la forme suivante, où les notions de chaînes anaphoriques et chaînes référentielles sont appréhendées plus en termes de types de liens qu'en termes de types de chaînes[3] :

---

[1] Corblin 1995, p. 167.
[2] Voir chapitre III § III.2.1.
[3] Dans ce schéma, les relations d'identité sont appelées « équi-référence » et les relations d'association « liens associatifs », voir Corblin 1995, p. 168.

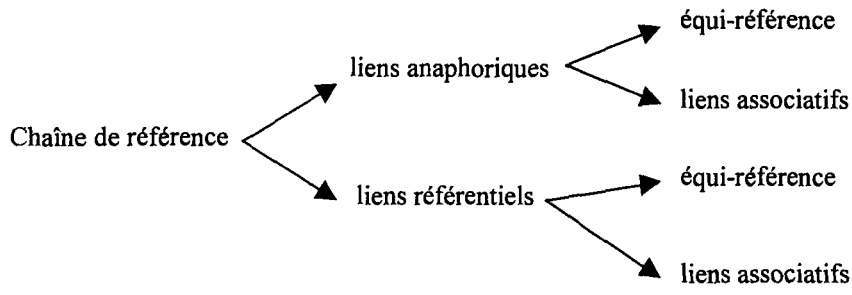


Figure 4 - Tableau des connexions référentielles en discours
Corblin 1995, p. 169

Les deux types de liens (d'identité et associatifs), actifs dans les deux types de chaînes (anaphoriques et référentielles), sont, dans un texte, constamment intriqués, de sorte que, typiquement, un texte en langue naturelle présente une configuration de type : « ...b ...x. ...a ....b ...x », où *x* note un élément variable et *a/b* des constantes[1].

Par la notion de chaîne de référence, reformulée en des termes linguistiques, il devient possible de montrer que la référence en discours se construit invariablement entre « des segments de nature formellement et interprétativement hétérogène », de formuler ensuite une règle de fonctionnement (« loi de la diversité formelle ») et de rattacher cette loi aux propriétés des textes eux-mêmes (notion de « plasticité textuelle »). L'approche linguistique permet également de spécifier le type d'unités linguistiques en cause dans les chaînes de référence (*infra*).

Si, d'un point de vue linguistique, la construction de la référence en discours peut être adéquatement représentée par le modèle des chaînes de référence, comment peut-on utiliser une telle formalisation pour déterminer la morphologie du descripteur en indexation ?

I.3.2 - DISCUSSION : CHAÎNE DE RÉFÉRENCE ET INDEXATION

À ce stade de notre recherche, nous avons pu définir l'indexation comme un processus qui doit permettre d'établir une relation référentielle, c'est-à-dire, compte tenu de la spécificité sémiotique des objets manipulés (des textes), de construire des thèmes de discours. Cette construction se réalise par le biais d'un parcours interdiscursif (mettant en jeu plusieurs textes). Nous nous sommes donné, dans le chapitre IV, les moyens de permettre une telle analyse interdiscursive : le regroupement des documents doit s'établir sur la base d'un *a priori* non formel.

L'indexation, comprise désormais au sens étroit de détermination de descripteurs, consiste donc, dans ce cadre, à sélectionner, dans une collection documentaire, les unités de discours susceptibles de permettre la construction de tels objets (des thèmes de discours).

---

[1] Corblin 1995, p. 185.

La représentation linguistique de la construction référentielle menée à travers le modèle des chaînes de référence nous fournit un cadre théorique dans lequel nous pouvons spécifier la morphologie du descripteur « en soi ». En effet, un descripteur peut être défini, grâce à ce cadre, en dehors de la problématique des langages documentaires[1]. Ce modèle, outre qu'il permet d'approcher le descripteur par rapport à son rôle dans les textes eux-mêmes, permet aussi de discuter la problématique classique du descripteur, approchée par le biais des langages documentaires.

En effet, le principe du langage documentaire (créer une liste de mots où la permanence linguistique, la permanence des formes, garantit la permanence référentielle, la permanence de la désignation) ne semble adéquat que pour les langages formels[2]. Or, les textes ne sont pas écrits dans des langages formels mais en langue naturelle. Pour peu que la spécificité textuelle des objets qu'elle manipule soit prise en compte, l'indexation ne peut recourir au principe des langages documentaires, qui sont alors, dans ce cas, de véritables « langages artificiels » appliqués à des objets « symboliques ».

La nécessité de prendre en compte, dans l'indexation, la spécificité des objets textuels (leur plasticité) repose sur les propriétés que l'on attribue au document. Si, comme nous l'avons montré dans le chapitre III, un document est choisi en fonction de sa rentabilité supposée, de ses possibilités d'usage et de détournement, l'indexation ne peut traiter des documents réduits à une seule de leurs interprétations possibles. L'indexation interprétative aboutirait, de façon caricaturale, à considérer qu'un document n'est susceptible d'apporter de réponse qu'à un seul type de question. Autrement dit, si le langage documentaire ne se justifie que dans le cas d'une indexation interprétative, une telle indexation, qui peut alors traiter les textes comme des objets symboliques, ne se donne plus, du coup, les moyens de « rentabiliser » un fonds documentaire. L'indexation conçue par le biais du langage documentaire aboutit donc à une contradiction entre la fonction dévolue au document en indexation (fonction multi-usages) et l'usage effectivement permis par l'indexation (mono-usage).

L'indexation doit donc être explicative : elle doit considérer les textes dans leur spécificité propre et pour cela les maintenir dans leur potentialité interprétative. Dès lors, le descripteur doit être défini par rapport aux textes eux-mêmes et non plus par rapport à un langage documentaire. C'est alors à un modèle de la référence discursive que se rattache la problématique du descripteur.

L'utilisation du modèle des chaînes de référence en vue de spécifier la morphologie du descripteur suppose que :

---

[1] Rappelons que, dans l'approche normative, un mot devient descripteur à partir du moment où il est intégré dans un langage documentaire, dans lequel il se voit attribuer une « stabilité référentielle ».

[2] Corblin 1987, p. 15 (déjà cité) : « Ce qui est souvent admis comme résultat idéal pour un automate dans ce domaine [exploitation documentaire de masses importantes de textes], c'est un texte en langue naturelle vidé des supports linguistiques des chaînes et présentant en lieu et place des identifieurs symboliques, des indices. Soit, à peu près : remplacer, en utilisant, des nombres ou des lettres, les expressions d'un texte par le symbole de ce qui est désigné. [...] Ce qui serait perdu c'est notamment ce qu'on pourrait appeler la plasticité des objets textuels naturels, propriété spécifique et étrangère aux systèmes formels ».

(i) d'un principe de réception textuelle (principe d'interprétation des référents discursifs), l'on puisse établir un principe de production lexicale (extraction d'unités de discours). Si l'interprétation de la référence textuelle se réalise par des moyens linguistiquement hétérogènes, que l'on peut représenter sous la forme de chaînes liant entre eux certains segments d'un texte, alors l'indexation, qui doit permettre la construction de la référence textuelle, doit extraire des textes les éléments de ces chaînes de référence. L'indexation se conçoit alors comme une opération guidée par une propriété linguistique qu'exhibent les textes eux-mêmes[1] ;

(ii) la notion de chaîne de référence doit pouvoir être valide non seulement à l'intérieur d'un seul discours, mais aussi entre plusieurs discours. Si, pour Chastain, les relations interdiscursives s'établissent typiquement au travers de chaînes référentielles, Corblin montre qu'une telle approche peut être prise en défaut : les chaînes référentielles se laissent caractériser selon lui, de façon plus adéquate, par la notion de chaîne communicative. Reste que, compte tenu de la nature de son projet (analyse des chaînes ou plutôt des liens anaphoriques), Corblin ne développe pas véritablement cet aspect des chaînes référentielles. Notre propre approche des chaînes référentielles restera sur ce point fidèle à Chastain[2] tout en prenant en compte les critères proposés par Corblin : les chaînes référentielles peuvent, nous semble-t-il, caractériser les liens interdiscursifs dans la mesure où notre notion de discours documentaire permet de prévoir une contrainte sur la construction des relations entre documents.

Si l'indexation est, nous semble-t-il, principalement concernée par les chaînes référentielles (interdiscursives selon Chastain), elle est nécessairement concernée aussi par les chaînes anaphoriques, du fait même que les deux types de chaîne s'entrecroisent selon un schéma de la forme[3] : un terme T_1 d'un discours D_1 est lié anaphoriquement par un terme T_2 du même discours D_1, lié référentiellement à un terme T_3 dans un discours D_2, etc.

Compte tenu de ce fonctionnement des chaînes de référence, l'indexation-extraction, telle que nous la concevons, devrait ne sélectionner des textes que des types d'unités susceptibles d'appartenir à la fois aux chaînes référentielles et aux chaînes anaphoriques. Élément d'une chaîne référentielle, le descripteur joue typiquement un rôle de « relais » entre discours ; considéré au sein d'une chaîne anaphorique, il joue alors un rôle de « pivot » dans un discours.

Pour simplement schématiser notre proposition, on considérera que, pour une chaîne de référence telle que notée en (3), seuls les termes T_2, T_3, T_4, pourront être des descripteurs susceptibles d'être pertinents, c'est-à-dire susceptibles de permettre la construction d'un thème de discours :

---

[1] Si l'on se place ici du point de vue de l'utilisateur, nous ne nous plaçons pas pour autant du point de vue de la recherche documentaire. L'hypothèse à partir de laquelle nous déterminons la morphologie du descripteur est une hypothèse de nature linguistique qui concerne l'interprétation textuelle et la formation des thèmes de discours. Pour étudier le descripteur en indexation, on n'est donc pas nécessairement obligé de faire d'autres types d'hypothèse (sur les besoins, ou encore les intérêts, des utilisateurs). De ce point de vue, notre appréhension de l'indexation reste non circulaire.

[2] Marandin [1988, p. 77, note 13] reste, sur les chaînes référentielles, fidèle lui aussi à l'esprit de Chastain et considère ce type de chaînes en termes de liens inter-textuels, même si son propos exploite essentiellement la notion de chaîne anaphorique.

[3] Repris de Chastain 1975, p. 214, déjà cité.

(3) Un terme T_1 d'un discours D_1 est lié anaphoriquement à un terme T_2 du même discours D_1, lié référentiellement à un terme T_3 dans un discours D_2, lié référentiellement à un terme T_4 dans un discours D_3, lié anaphoriquement par un terme T_5 du même discours D_3, etc.

Autrement dit, le descripteur serait un type de terme singulier pourvu de propriétés lui permettant d'appartenir à la fois aux chaînes référentielles et aux chaînes anaphoriques.

Pour tenter de spécifier la morphologie du descripteur sous l'angle de cette contrainte, nous examinerons d'abord les types d'unité linguistique mis en jeu dans les chaînes référentielles, car c'est d'abord elles qui au premier chef nous intéressent ; nous verrons ensuite si l'un de ces types d'unité peut également constituer un élément d'une chaîne anaphorique.

I.3.3 - LES UNITÉS LINGUISTIQUES EN JEU DANS LA CONSTRUCTION DES CHAÎNES DE RÉFÉRENCE

Si, pour Chastain, les types de chaînes de référence ne se distinguent pas crucialement par le type de termes singuliers qu'elles mettent en cause, Corblin[1] établit, lui, dans une perspective d'analyse linguistique, les distinctions suivantes :

- une chaîne anaphorique s'établit à travers les types d'unités linguistiques suivants : pronoms, démonstratifs, descriptions définies (complètes et incomplètes) ;

- une chaîne référentielle met essentiellement en jeu des noms propres et des descriptions définies complètes.

Compte tenu du fonctionnement attendu du descripteur (élément d'une chaîne anaphorique et d'une chaîne référentielle), on dispose de deux candidats-descripteurs aux chances « inégales » :

- Les noms propres ne semblent pouvoir être utilisés que dans les chaînes référentielles. Ils ne remplissent qu'une des conditions requises pour être descripteurs. Cependant, comme nous l'avons vu précédemment, les pratiques d'indexation recourent régulièrement à des formes « nom propre » utilisées comme descripteurs. Quelles sont les propriétés du nom propre qui empêchent, dans le cadre de notre approche, de le considérer comme un bon candidat-descripteur ? Peut-on expliquer pourquoi, dans l'indexation classique que l'on a pu qualifier d'interprétative, ce type d'unité linguistique puisse fonctionner comme descripteur ?

- Les descriptions définies complètes apparaissent, elles, à la fois dans les chaînes anaphoriques et dans les chaînes référentielles : elles semblent constituer, à première vue, de bons candidats descripteurs. En ce sens, nous nous intéresserons de près à leur fonctionnement logique et à leurs propriétés linguistiques.

---

[1] Corblin 1995, p. 151-195.

Avant d'entamer une discussion plus précise sur les « chances » respectives des deux candidats-descripteurs qui se présentent dans notre cadre d'analyse de l'indexation au travers des chaînes de référence, nous présentons succinctement la notion de « description définie ». Nous renvoyons, pour une première approche logique du nom propre, à notre présentation[1] de la notion de rigidité établie par Kripke : c'est sur elle que s'appuie Corblin pour représenter le rôle du nom propre dans les chaînes référentielles.

Qu'elles soient « complètes » ou « incomplètes[2] », les descriptions définies, ou, pour reprendre les termes de Corblin, les groupes nominaux[3] définis, se définissent de la même façon. Dans les deux cas, il s'agit d'une séquence morphologique de type *le* + *N* à *interpréter* comme désignateur « sur la base du signalement qui suit *le*, c'est-à-dire qu'il faut associer au groupe nominal un domaine d'interprétation où la description préfixée soit singularisante ». Les GN définis ont en effet pour rôle de repérer un individu dans un ensemble, c'est-à-dire d'identifier un objet parmi un ensemble d'objets d'une même classe[4]. Par exemple, le groupe nominal « le président des USA » renvoie à la classe des présidents américains : c'est le contexte d'apparition du GN qui permet d'isoler un seul président parmi les différents qui se sont succédé.

La propriété des GN définis consiste à établir cette double référence et à une classe et à un objet de cette classe. L'identification d'un objet parmi une classe s'établit à la fois sur la base de la signification des éléments lexicaux qui constituent le GN et sur la base du contexte où apparaît le GN.

Ce qui différencie les GN définis complets des GN définis incomplets réside dans le type de contexte sollicité : nous y reviendrons. Essayons tout d'abord de préciser comment s'effectue le mécanisme d'identification d'un élément et le rôle respectif de la signification lexicale et du contexte discursif, sachant que la construction du référent discursif nécessite que soient fixés à la fois un domaine d'interprétation et un critère de sélection : « Pour un groupe nominal défini, il s'agit toujours de déterminer sur la base de son contenu descriptif un domaine d'interprétation et un critère de sélection tel que ce critère ne s'applique qu'à un élément du domaine.[5] »

A - Identification d'un élément d'une classe

Pour capter la dimension référentielle des unités lexicales hors discours, on a précédemment proposé[6] de la représenter par le biais de la notion de « référence virtuelle » proposée par Milner. Cette notion qui caractérise les éléments lexicaux du GN (« président » dans l'exemple « le président des USA ») et qui constitue leur « sens » dans le modèle de Milner, correspond en partie[7] à la notion de « contenu descriptif » que Corblin reprend de Chastain. Les deux notions captent un même principe linguistique de découpage du « domaine d'interprétation » d'un GN : « Les références définies sont réalisées au moyen du contenu descriptif du GN utilisé

---

[1] Voir chapitre III, § III.2.
[2] La distinction est ici exprimée en des termes logiques ; les distinctions établies sur un plan linguistique sont quelque peu différentes : nous y revenons ci-après.
[3] Abrégé GN désormais.
[4] Corblin 1995, p. 163 : « Il est seulement requis pour les GN définis que le contexte fixe un domaine de référence où le GN sera capable d'isoler un individu du reste ».
[5] *Ibid.*, p. 192.
[6] Chapitre III, § II.2.
[7] Voir ci-après, point C de ce paragraphe.

comme signalement discriminant sur un domaine d'objets. Il s'agit donc, lorsqu'un GN défini est utilisé, d'identifier un tel domaine dans le contexte. [...] C'est donc le contenu linguistique du GN qui guide le processus d'identification du domaine et par là, la nature de l'emprunt contextuel requis.[1] »

Pour illustrer ce fonctionnement des descriptions définies, prenons le cas simple[2] du groupe nominal « la voiture » dans la phrase : « La voiture roulait vite ». L'usage du déterminant *la* signifie qu'il y a saisie d'un individu déterminé (le défini *la* fonctionne comme « critère de sélection ») ; cette saisie est possible grâce à un signalement singularisant donné par la signification lexicale du *N*, c'est-à-dire par la référence virtuelle de « voiture » (qui correspond à la classe d'objets virtuellement « désignables » par « voiture »). Reste à identifier l'individu réellement singularisé par le groupe nominal « la voiture » : c'est là encore le contenu lexical du *N* « voiture » qui permet de repérer, dans le contexte extra-phrastique où apparaît le groupe nominal « la voiture », une autre occurrence qui va, elle, pouvoir spécifier le référent en cause. Cette autre occurrence peut être de nature très différente ; on peut avoir, par exemple :

« *une voiture* nous a dépassé »
« Il y a eu *un accident* »
« J'ai croisé *les Dupont* », etc.

On retrouve ici les notions de contrainte lexicale (c'est la signification lexicale des éléments d'un groupe nominal qui guide la construction d'une classe d'objets) et de sous-détermination référentielle (c'est en discours qu'un domaine d'interprétation va pouvoir être identifié et un objet isolé) que nous avions précédemment dégagées. C'est cette sous-détermination qui permet la multiplicité de saisies possibles, c'est-à-dire la possibilité d'identifier des objets différents appartenant à une même classe ; c'est en cela que les descriptions définies peuvent constituer des chaînes de référence.

B - Différence entre types de description définie

L'identification d'un objet correspond, sur un plan linguistique, à la « saturation » d'un groupe nominal (fixation d'une valeur). Cette saturation peut s'effectuer de deux façons différentes en discours, le principe étant qu'un domaine ou encore qu'un « point de référence » soit fixé[3] :

- soit « une mention fournit effectivement comme source un individu de type *N*[4] », comme dans l'exemple suivant : « Ils entrèrent dans *la maison*. Dans *la cuisine*, ils virent un homme qui lisait sous *la lampe* ». Les GN définis « la cuisine » et « la lampe » s'interprètent par le biais de la mention « la maison » qui fournit le domaine d'interprétation des mentions ultérieures ;

- soit « une mention fournit un domaine où nos connaissances du monde nous permettent de savoir qu'il n'y a qu'un seul *N*[5] ». Dans l'exemple suivant, c'est nos connaissances de la langue jointes à notre connaissance du monde

[1] Corblin 1995, respectivement p. 163 et p. 164.
[2] Repris de Corblin 1995, p. 62.
[3] *Ibid.*, p. 165-166 : « Le défini exige de son contexte la figuration de points de référence constituant un domaine où son contenu descriptif soit singularisant ».
[4] *Ibid.*, p. 165.
[5] *Id.*

(monogamie) qui nous permettent d'identifier un seul individu : « J'ai beaucoup aimé la soirée d'hier. *Le mari de Jeanne* est formidable ».

Dans le premier cas, où les GN définis sont saturés par le biais d'un point de référence nommé dans le texte, on parlera de *descriptions définies incomplètes* ; dans le second, où le point de référence saturant le GN défini est hérité du contexte de communication, on parlera de *descriptions définies complètes*.

Dans les deux cas, l'identité référentielle est *calculée* entre éléments textuels formellement hétérogènes. Le renvoi à un univers d'objets n'est pas direct, il passe par le « renvoi d'une forme à une autre forme et un lieu de discours[1] ». Ces suites de renvoi reposent sur la propriété des GN définis, sur leur possibilité de désigner à la fois une classe et un élément singulier de cette classe. La distinction entre types de GN définis (saturés ou pas) ou types de descriptions définies (complètes ou pas) réside dans la façon dont se réalise la fonction singularisante du GN (le repérage d'un référent unique) : uniquement sur des « bases communicatives » ou aussi sur des bases linguistiques.

C - Référence discursive des descriptions définies

En cours de chaîne de référence, au fur et à mesure qu'un objet de discours se construit dans une chaîne, le contenu descriptif lui-même des termes de la chaîne se modifie pour parvenir à isoler un référent discursif qui pourra être vu *in fine* comme un thème de discours.

Rappelons que, dans l'approche proposée par Marandin, un thème de discours se laisse représenter comme un « individu composite », c'est-à-dire comme une association entre un « terme textuel » et un « contenu descriptif ». Le « terme textuel » correspond à la matérialisation lexicale de l'objet de discours d'une chaîne de référence (« un individu relativement à un texte[2] »). Le « contenu descriptif » d'un thème se donne sous la forme d'un « agrégat » de discours, « un agrégat subsumant d'autres individus dans leurs interrelations, telles qu'elles sont introduites dans les énoncés, reconstruites dans la compréhension et constitutives d'une interprétation ».

En cela, la notion de contenu descriptif que Marandin reprend de Chastain se distingue de la notion de référence virtuelle. Au fur et à mesure des discours, les termes singuliers parviennent à spécifier un objet de discours spécifique grâce aux différentes propriétés que leur a attribué la chaîne des phrases à laquelle ils appartiennent. Pour Chastain, les termes singuliers d'une chaîne de référence acquièrent un pouvoir de singularisation qui ne repose plus uniquement sur leur dénotation mais aussi sur leur référence discursive : « Thus, the descriptive content of a singular term in a discourse is, roughly speaking, what sort of thing the discourse says the referent is supposed to be, what properties it is supposed to have, what sorts of other things is supposed to be related to and in what ways, and so on. *The term not only purports to refer but purports to refer to a thing of a specified kind.*[3] »

---

[1] Corblin 1995, p. 174.
[2] Marandin 1988, p. 82.
[3] Chastain 1975, p. 230. (C'est nous qui soulignons).

C'est en vertu de cette caractéristique du contenu descriptif des termes singuliers dans les chaînes de référence que nous ferons l'hypothèse que, dans les chaînes en jeu en indexation, ce ne sont pas seulement les descriptions définies complètes qui peuvent constituer de bons candidats-descripteurs. D'autres types de descriptions définies, qui ne sont pas distinguées d'un point de vue logique, sont, nous semble-t-il, à prendre en considération : entre les descriptions définies incomplètes (comme « l'étoile ») et les descriptions définies complètes (comme « l'étoile du Berger »), nous aurons à considérer des descriptions définies du type « l'étoile des mers » ou « la rose des vents », qui permettent, sans recourir à la mention effective d'une source antérieure, d'isoler en discours à la fois une classe et un individu sur la base de leur contenu descriptif, relatif ici à des domaines d'usage constitués d'autres discours. En ce sens, notre exploration des candidats-descripteurs considérera les descriptions définies dans leur ensemble, sans tenir compte de la distinction, proprement logique, entre les descriptions définies complètes et incomplètes.

1.3.4- CONCLUSIONS INTERMÉDIAIRES

Avant de poser un cadre théorique (celui des chaînes de référence) permettant d'étudier la morphologie du descripteur, nous avons jugé nécessaire de situer notre problématique du descripteur comme unité de discours et celle, concomitante, de l'indexation comme extraction. Il nous importait de montrer que la problématique du descripteur comme unité de discours était loin d'être triviale et exigeait une remise en cause des modes d'appréhension classiques. Sur ce point, nous avons cherché à montrer que notre approche de l'indexation comme processus de détermination des thèmes discursifs pouvait rendre compte à la fois de l'indexation-assignation (dite « classique ») et de l'indexation-extraction que nous proposons. Si nous défendons la nécessité de pratiquer le second type d'indexation, nous avons pu comprendre la prédominance des formes classiques d'indexation ; des contraintes, en partie de nature matérielle, peuvent expliquer la pratique massive de l'indexation que nous avons qualifiée d'interprétative.

Désormais, l'évolution technologique aidant, l'indexation peut se dégager des contraintes matérielles et envisager de prendre un nouveau tournant. Notre contribution s'inscrit dans le cadre de cette évolution. Nous avons proposé que, dans une approche de l'indexation qui intègre la notion de chaînes de référence, le descripteur soit un terme singulier susceptible d'appartenir à une chaîne référentielle : à ce titre, un descripteur peut être, en adoptant les propositions de Corblin, soit un nom propre soit une description définie complète. Cependant, compte tenu de l'intrication des chaînes de référence en discours, il est apparu que le descripteur devait aussi pouvoir appartenir à une chaîne anaphorique : à ce titre, le nom propre apparaît comme un moins bon candidat-descripteur ; il est cependant régulièrement utilisé comme unité d'indexation. Parallèlement, d'autres types de descriptions définies (que les descriptions définies complètes) pourraient constituer des descripteurs ; mais elles ne sont pas toujours distinguées par la seule approche logique.

Pour poursuivre notre examen des candidats-descripteurs en nous focalisant sur ces deux types d'unité, nous procéderons successivement à deux types d'analyse, l'une menée d'un point de vue logique (II), l'autre d'un point de vue linguistique (III).

II - Approche logique du descripteur

Notre approche du descripteur comme élément de chaîne de référence nous désigne deux candidats possibles (les noms propres et les descriptions définies) que nous examinerons d'abord sous l'angle du rôle « logique » que l'on peut attribuer au descripteur.

II.1 - Rôle « logique » du descripteur

Lorsqu'il est élément d'une chaîne de référence, le descripteur est un terme singulier dont le « contenu descriptif » (ce que le ou les discours dit/disent de son référent) lui permet de fonctionner comme un « terme textuel », pour reprendre la terminologie proposée par Marandin[1].

Rappelons que Marandin examine, dans son étude, le thème de discours lui-même, alors que notre approche s'intéresse aux « segments textuels » d'une chaîne qui permettent l'interprétation et la nomination du thème par un lecteur, un utilisateur d'un système d'information. Cependant, du fait que la construction d'un thème discursif peut se comprendre comme l'interruption d'une lecture, ou encore comme la rupture d'une chaîne de référence, il n'y a pas toujours de différence significative, nous semble-t-il, entre les segments textuels que l'on juge terminaux et ceux que l'on juge intermédiaires. Les segments textuels intermédiaires se chargent, au fil du texte, de propriétés héritées du texte lui-même ; c'est de ce point de vue-là que certains types de segments textuels (et notamment les descriptions définies) peuvent être considérés, nous semble-t-il, comme des termes textuels. Reste que, du point de vue de l'interprétation du thème discursif, une distinction majeure[2] est à maintenir entre les unités susceptibles d'être « nomination » de thème[3] et celles qui ne le peuvent pas.

C'est donc uniquement du point de vue de leur fonctionnement en discours (et non du point de vue de leur interprétation discursive) que les descripteurs nous semblent pouvoir être représentés par la notion de « terme textuel » proposée par Marandin : en cela, nous ferons un usage plus large de la notion proposée par l'auteur.

En tant que terme textuel, le descripteur pointe sur un objet (un objet de discours, un référent) par le biais d'un ensemble de propriétés relatives à sa classe (notamment l'ensemble des discours portant sur d'autres objets de la même classe). Autrement dit, le descripteur pose une relation dialectique entre classe et objet, qui s'apparente, d'un certain point de vue, à une prédication, si l'on entend par prédication la mise en rapport de deux entités indépendantes, un objet et une description[4]. À ce titre, on peut dire que le rôle « logique » du descripteur consiste à établir une double désignation, à la fois à une classe et à un objet de cette classe : c'est sous l'angle de cette contrainte que nous examinerons le fonctionnement du nom propre, d'une part, et celui des descriptions définies, d'autre part.

---

[1] Marandin 1988.
[2] La distinction est nette dans Marandin [1988] où il montre que la construction du thème en discours met en jeu plusieurs éléments du texte, de différentes natures : linguistique (rôle des transitions temporelles dans les textes) et cognitive (rôle des « schèmes cognitifs » qui permettent de repérer par exemple les rapports de causalité dans un texte).
[3] Marandin [1988] propose d'appeler de telles unités des « SN-fermoirs » dans le cas du thème configuré.
[4] *Ibid.*, p. 75.

II.2 - Examen des candidats-descripteurs

Comme le rappelle Kleiber [1981], description définie et nom propre constituent, parmi l'ensemble des termes singuliers, une sous-classe spécifique d'unités qui présentent la particularité de pouvoir s'employer « aussi bien en l'absence qu'en présence du référent[1] ». En cela, ces unités s'opposent à ce que Kleiber nomme les « indicateurs », classe qui recouvre les pronoms personnels, les pronoms démonstratifs et les descriptions démonstratives[2].

Si description définie et nom propre se ressemblent de ce point de vue, ils diffèrent dans la façon dont ils réalisent la singularisation de leur référent.

II.2.1 - FONCTIONNEMENT LOGIQUE DU NOM PROPRE

La notion de nom propre est envisagée ici, comme dans d'autres passages de cette recherche, uniquement sous un angle logique : elle correspond à la notion de « désignateur rigide » proposée par Kripke et couvre uniquement les cas où le nom propre est employé sans article[3].

Sous cet angle, la rigidité désignative des noms propres pourrait *a priori* en faire des candidats-descripteurs idéals : à une identité linguistique correspond une identité référentielle et l'indexation ne rencontre aucune des ambiguïtés référentielles à laquelle elle se trouve régulièrement confrontée. Cependant, l'examen du nom propre comme candidat-descripteur dans une chaîne de référence montre qu'une telle évidence mérite d'être nuancée : le nom propre ne constitue pas, sauf exception, une classe (A), et ne peut donc que difficilement participer à la construction de l'objet de discours (B). Cependant, ce n'est pas pour autant que les propriétés du nom propre ne peuvent être exploitées en indexation : mais le nom propre n'a plus alors le statut de descripteur ; un autre statut devra être défini[4].

A - Désignation simple à un individu : pas de construction de classe

D'un point de vue formel, la notion de « rigidité » établie par Kripke peut être rapprochée d'une « fonction constante », qui, à chaque élément x d'un domaine (un « monde possible ») associe une constante individuelle a[5]. Si cette constante individuelle reste stable d'un monde possible à un autre, en revanche, les propriétés de l'individu qu'elle désigne peuvent, elles, varier. Il semble donc que si le nom propre désigne directement un individu sans passer par l'examen de ses propriétés, il ne peut signaler l'appartenance d'un individu à une classe[6].

---

[1] Kleiber 1981, p. 309.
[2] *Id.*
[3] C'est ce type d'objet qu'étudient les logiciens, Gary-Prieur 1994, p. 16 : « Les logiciens, eux, ne s'intéressent (jusqu'à présent au moins) qu'au nom propre employé sans article en position référentielle ».
[4] Nous esquissons quelques pistes en II.3.
[5] Nef 1991, p. 103.
[6] La notion de classe est entendue ici au sens logique d'« ensemble de propriétés », et, sur ce point, elle n'est pas équivalente à celle de collection d'individus. Si l'on considère les cas où un nom propre est porté par des individus différents, on dira qu'il renvoie alors à une collection et non à une classe, sur ce point voir Gary-Prieur 1994, notamment p. 98 et p. 166.

Ne pouvant lier une description à un objet, le nom propre ne peut donc établir de prédication, de relations entre objets de même type. À ce titre et en suivant Fauconnier, on dira que le nom propre désigne une valeur mais qu'il n'a pas de rôle (« prédicatif »), sauf « dans certaines conditions pragmatiquement appropriées[1] ».

Ce sont précisément ces conditions particulières que, d'un point de vue singulièrement pragmatique, l'indexation « classique » réalise et exploite.

B - Affectation d'un rôle à un nom propre : constitution d'une classe

Rien n'empêche d'attribuer, de façon arbitraire, un rôle à un nom propre. L'exemple de Fauconnier est sur ce point particulièrement éclairant[2]. Imaginons des voisins qui appelleraient systématiquement leur chien *Médor* (ils n'ont qu'un chien à la fois). Dès lors, le nom propre *Médor* désigne non seulement une valeur, un individu, mais aussi un rôle, ou encore un prédicat (en l'occurrence « être le chien du voisin »), qui constitue une classe (ici un singleton).

Affecter un rôle à un nom propre revient à lui attacher une description définie et donc à le « remotiver ». En effet, on sait que lors du « baptême initial » du nom propre, une description définie lui est affectée, pour être aussitôt oubliée, puisque c'est précisément cet oubli qui permet au nom propre d'être rigide, c'est-à-dire d'être imperméable aux différentes propriétés que pourrait lui attribuer un contexte[3]. L'affectation d'un rôle, d'une nouvelle description définie, réanime une nouvelle chaîne « communicative » dans les termes de Kripke, « référentielle » dans ceux de Chastain : il y a donc, nécessairement, rupture avec la chaîne de référence précédente, rupture nécessaire pour établir une nouvelle stabilité référentielle. Cependant, contrairement aux descriptions définies, le lien établi entre rôle et valeur dans le nom propre n'est pas linguistiquement contraint. En cela, le nom propre, même affecté d'un rôle, ne pourra fonctionner dans une chaîne anaphorique : le type de classe qu'il constitue présente en effet des propriétés particulières.

Avant d'examiner ces particularités, intéressons-nous à l'usage documentaire qu'il est fait du nom propre pourvu de rôle dans les pratiques d'indexation classique.

B1 - Usage du nom propre en indexation

Il nous semble que les pratiques d'indexation classique, de type interprétatif, exploitent le nom propre comme descripteur en lui affectant plus ou moins implicitement un rôle. C'est clairement le cas dans les systèmes documentaires qui structurent leurs données en distinguant, par exemple, les rubriques « auteur » et « personnes citées » susceptibles d'accueillir des noms propres : l'indexation par nom propre pourvu de valeur est ici évidente, cependant elle n'est pas toujours, nous semble-t-il, interprétée de façon adéquate[4].

---

[1] Fauconnier 1984, p. 92.
[2] *Ibid.*, p. 90-91.
[3] Voir, sur ce point, chapitre III § III.2.
[4] Si les professionnels s'accordent tous à remarquer qu'en matière de recherche d'informations, c'est l'interrogation par « auteur » qui fonctionne le mieux, ils ne relient pas souvent le succès de ce type d'interrogation avec l'unité linguistique qui en est la cause (cf. Calenge 1994 : « On savait bien que, pour les usagers, le recherche par auteur était la seule à peu près fiable, et que les autres s'apparentaient à un butinage plus ou moins expert »). Sur

Il est des cas où l'affectation d'un rôle à un nom propre reste complètement implicite. Notre enquête[1] exhibe certains de ces emplois implicites du nom propre pourvu de rôle. Ainsi, un article du *Monde*[2] consacré aux redressements judiciaires des sociétés appartenant à Bernard Tapie est-il majoritairement[3] indexé par le nom propre « Bernard Tapie ». Certes, les sociétés appartenant à Bernard Tapie portent, pour certaines d'entre elles, la marque linguistique de leur entrepreneur[4]. Reste, sur les plans linguistique et référentiel, une différence essentielle entre l'indexation par le nom propre « Bernard Tapie » et l'indexation par le nom propre « Bernard Tapie Finances » (par exemple), voir ci-après B2. En outre, ce type d'indexation, en brisant les chaînes référentielles constituées par le nom propre Bernard Tapie (dépourvu, lui, de rôle), n'est pas sans créer des ambiguïtés interprétatives, la classe que constitue le nom propre pourvue de rôle étant d'une nature particulière (B3).

B2 - L'indexation par nom propre pourvu de rôle : principe de la métonymie

En reprenant le cadre de l'analyse proposée par Bonhomme[5], on dira que, dans le cas où le nom propre est affecté d'un rôle (par exemple « être entrepreneur »), l'indexation emprunte le principe de la métonymie, alors que, dans le second cas, où le nom propre n'a qu'une valeur et pas de rôle, l'indexation recourt au principe de la synecdoque. Les mécanismes référentiels en jeu ne sont pas les mêmes.

En effet, à la fonction « individualisante » de la métonymie s'oppose la fonction « focalisante » de la synecdoque : si la métonymie opère un déplacement référentiel (vers un individu), la synecdoque ne réalise qu'un resserrement référentiel (sur un individu).

Bonhomme décrit ainsi le transfert métonymique du nom commun au nom propre : « La métonymie possède une fonction individualisante dans toutes les occurrences où le transfert tropique substitue un nom propre à un nom commun, autrement dit une dénotation référentielle à une dénotation pluri-référentielle.[6] »

Dans les termes de Kripke, le mécanisme décrit par Bonhomme reviendrait à une opération de recouvrement de l'existence *in dicto* par l'existence *in re*[7]. Le rôle affecté au nom propre ou, plus exactement, les modalités discursives attribuées, restent en effet seconds : « Que la métonymie soit situative [cas des séries troponymiques] ou actancielle [cas des séries anthroponymiques], l'individualisation provient du rôle-clef que joue le facteur référentiel, invariablement déterminatif, à l'intérieur du nom propre tropique, cela aux dépens

---

ce point encore, on peut remarquer combien il est difficile d'analyser les faits d'indexation, ou plus largement les faits documentaires, quand on ne sort pas du référentiel documentaire : la notion d'« auteur » n'est pas ici suffisante pour comprendre le succès de ce type d'interrogation.

[1] Présentée en annexe 1.

[2] *Le Monde* du 01/12/1994, p. 24.

[3] 4 cas sur 7.

[4] Bernard Tapie Finances, Groupe Bernard Tapie, la Financière Immobilière Bernard Tapie mais aussi La Vie claire, Terraillon, Testut, Scaime, Alain Colas Tahiti.

[5] Bonhomme 1987.

[6] *Ibid.*, p. 126.

[7] Voir chapitre III, § III.2.1.

du facteur sémantique (générique par nature), toujours important dans le nom commun, qui constituerait la dénotation attendue.[1] »

C'est ainsi que l'usage du nom propre pourvu de rôle en lieu et place de « descripteur thématique » (de descriptions définies) peut conduire à une indistinction référentielle, autrement dit à des ambiguïtés[2]. Nous avions pu, sur ce point, relever le risque d'ambiguïté que comportait l'indexation d'une interview de Simone Veil sur le sida par le nom propre « Simone Veil[3] ». Pariant, d'une certaine façon, sur la rigidité désignative du nom propre, l'indexation interprétative, en recourant au désignateur rigide pourvu de rôle, effectue implicitement une remotivation du nom propre ; mais, ce faisant, le nom propre perd tous les atouts de sa rigidité et introduit plus de risques d'ambiguïté.

Avec la synecdoque, on retrouve cette même indétermination référentielle, mais pas sous la forme d'une ambiguïté. En effet, la synecdoque, sans affecter le référent, le présente sous un point de vue particulier – « focal » –, qui se signale de la façon suivante[4] :

- d'une part, « une lourde charge référentielle [pèse] sur l'unité qui assume la dénotation de l'ensemble » ;
- d'autre part, se réalise « un resserrement dénotatif maximal de l'ensemble, perçu à travers une unité-type ».

Le rapport synecdochique, parce qu'il préserve, contrairement au transfert métonymique, l'espace discursif en premier plan (même sous une forme réduite), semble donc plus apte, de ce point de vue, à construire un descripteur : « La focalisation synecdochique ne se contente pas de parcelliser la dénotation ; elle provoque une violente compression dénotative de l'entité visée par l'énoncé, à l'issue de laquelle cette entité est non seulement fractionnée mais réduite au strict minimum qui permet de la reconnaître encore. [...] C'est dans ce *hiatus entre la restriction apparente du discours et l'étendue de son objet* que se situent le force et l'intérêt de la fonction focalisante.[5] »

L'utilisation synecdochique du nom propre, si elle paraît plus adaptée à l'indexation, correspond au cas où le nom propre, dépourvu de rôle, ne crée pas de classe.

Autrement dit, le nom propre parvient difficilement à conjoindre les propriétés attendues du descripteur dans le cadre des chaînes de référence en indexation : soit il peut fonctionner comme un terme textuel parce qu'il condense un discours (utilisation synecdochique) mais alors il ne constitue pas de classe ; soit il constitue une classe mais celle-ci passe par une rupture de chaîne référentielle (nom propre pourvu de valeur). Reste qu'il nous semble possible de déterminer un usage documentaire du nom propre qui ne soit pas celui du descripteur proprement dit (II.3).

---

[1] Bonhomme 1987, p. 127.
[2] *Ibid.*, p. 161, sur la métonymie auteur/œuvre : « Le trope métonymique peut causer une indistinction entre un individu et une œuvre artistique qui lui est intimement associée ».
[3] Voir chapitre III, § III.2.2.
[4] Bonhomme 1987, p. 167-168.
[5] *Ibid.*, p. 173 (c'est nous qui soulignons).

B3 - Nature de la classe construite par le nom propre pourvu de rôle

Comme nous l'avons précédemment mentionné, attribuer un rôle à un nom propre revient, dans les termes de Kripke, à re-baptiser un nom propre par le biais d'une description définie, mais aussi à préserver le lien entre les deux types d'unités : le nom propre semble donc fonctionner, dans ces cas, sur le même mode prédicatif que le nom commun (ou encore les descriptions définies). Cependant, il importe de noter que le fonctionnement prédicatif en jeu dans l'un et l'autre cas diffère sensiblement : la classe construite par le nom propre pourvu d'un rôle est constituée d'une partie des propriétés d'un objet singulier et non de propriétés générales communes à un ensemble d'objets[1]. En ce sens, deux occurrences d'un même nom propre, dont l'une est un nom propre pourvu de rôle et l'autre pas, ne peuvent appartenir à la même chaîne référentielle : ces deux mentions ne constituent pas des « ressaisies » différentes d'un même référent discursif, il s'agit au contraire de saisies différentes. L'exemple suivant[2] illustre ce cas de figure :

(5) Gide n'est pas né *Gide*

Le nom propre Gide en position sujet désigne l'individu singulier auquel l'a associé un « acte de baptême », dans les termes de Kripke. Le nom propre Gide en position attribut « prédique certaines propriétés de cet individu[3] », propriétés issues d'autres discours, d'une chaîne « communicative » différente de celle qui a permis de « baptiser » Gide. La classe construite par la seconde occurrence de Gide correspond au rôle qu'il se voit attribuer ; ce rôle peut être par exemple « être un écrivain reconnu ». La classe ainsi constituée ne permet pas au second nom propre de se comporter comme une description définie ou encore comme un nom commun[4] : la seconde occurrence de Gide ne semble pas en effet pouvoir être interprétée comme élément de chaîne anaphorique.

Il y a, nous semble-t-il, une différence d'interprétation entre (6) et (7) ci-dessous :
(6) Gide ne serait plus *Gide* si...
(7) Gide ne serait plus *lui-même* si...

Comme le fait remarquer Gary-Prieur[5], en (7), c'est l'identification entre x_i (Gide) et x_i (lui-même) qui est niée, alors que, dans (6), c'est l'attribution à x_i (Gide) d'un ensemble de propriétés (son rôle *Gide*) qui est niée.

La nature de la classe constituée par le nom propre pourvu de rôle nous indique que le nom propre ne constitue pas un bon candidat descripteur. Si, dans certaines conditions d'emploi, il parvient à constituer une classe, au sens d'ensemble de propriétés, le type de classe créée ne lui donne pas la possibilité d'apparaître dans une chaîne anaphorique.

Cette analyse du fonctionnement logique du nom propre nous a paru nécessaire au moins pour deux raisons. D'une part, le nom propre, s'il ne constitue pas, dans le

---

[1] Gary-Prieur 1994, p. 94 : « Alors qu'un nom commun attribue à un objet des propriétés générales, un nom propre caractérise toujours par le biais d'une identification au référent initial ».
[2] Repris de Gary-Prieur 1994, p. 77.
[3] *Id.*
[4] Comme le fait remarquer Fauconnier [1984, p. 93], un rôle ne peut être identifié par le nom de l'individu qui se trouve le remplir.
[5] *Ibid.*, p. 79.

cadre d'une indexation explicative, un descripteur à proprement parler, n'en révèle pas moins des propriétés qui peuvent lui faire tenir un rôle en indexation (II.3) ; d'autre part, cette analyse nous fournit quelques éléments pour expliciter le recours aux noms propres dans les pratiques d'indexation courantes.

En effet, l'utilisation du nom propre comme descripteur trouve son entière justification dans le cadre d'une indexation interprétative :

- *le nom propre permet, dans un emploi synecdochique (dépourvu de rôle), de fonctionner comme un « terme textuel », susceptible de condenser un discours. À ce titre, Marandin*[1] *remarque qu'une des façons les plus « immédiates » de nommer un thème de discours consiste à recourir à un nom propre. On retrouve ici la caractéristique de l'indexation classique (interprétative) qui est de proposer des thèmes de discours, des nominations de thèmes, et non des éléments permettant de les construire et de les nommer ;*

- *dans son emploi métonymique, le nom propre pourvu d'une valeur permet de redonner à un individu un « contenu », de revivifier certaines propriétés d'un individu. Il s'agit, là encore, d'une caractéristique typique de l'indexation interprétative qui offre une représentation volontairement partielle (et dès lors figée) d'un individu ; le nom propre pourvu d'un rôle ne présente qu'un sous-ensemble de propriétés singulières issues d'un ensemble de discours.*

Si l'emploi du nom propre comme descripteur est tout à fait compréhensible dans la perspective d'une indexation interprétative, il n'en reste pas moins que cette pratique d'indexation doit pouvoir : (i) d'une part, signaler de façon systématique les rôles attribués aux noms propres (en étendant par exemple le principe des rubriques existantes comme « auteur » ou « personne citée ») ; (ii) d'autre part, distinguer ces emplois de ceux où les noms propres sont dépourvus de rôle, faute de quoi la rupture d'une chaîne référentielle ou encore le passage d'une chaîne référentielle à l'autre risque de ne plus être perceptible : peuvent apparaître les risques d'ambiguïté que nous avons signalés à partir de quelques exemples[2].

II.2.2 - FONCTIONNEMENT LOGIQUE DES DESCRIPTIONS DÉFINIES

Le fonctionnement logique des descriptions définies s'oppose à celui du nom propre (A) en ce qu'elles commencent par spécifier un rôle (une classe d'objets pourvus du même rôle) qui permet ensuite de désigner une valeur (un objet singulier de cette classe). Cependant (B), il est des types de descriptions définies qui semblent, tout comme le nom propre, établir une relation référentielle directe avec l'objet sans nécessiter le passage par une classe : la distinction rôle/valeur n'y semble plus pertinente. L'examen de ces cas conduit à préciser la notion de classe : s'il y a toujours construction de classe dans le cas des descriptions définies, cette classe peut n'être que virtuelle (C).

A - L'ordre rôle/valeur dans les descriptions définies

Dire que les descriptions définies « désignent d'abord des rôles et seulement ensuite des valeurs par identification[3] » revient à dire que la relation à un objet n'est pas,

---

[1] Marandin 1988, p. 79.
[2] *Supra*, les cas d'indexations par « Bernard Tapie » ou par « Simone Veil ».
[3] Fauconnier 1984, p. 197.

avec ce type d'unité, directe. Comme le met en valeur Corblin, la relation indirecte à l'objet repose sur le mode de fonctionnement interprétatif spécifique des descriptions définies : « Les groupes nominaux définis semblent voués à la mise en relation *indirecte* en raison même du fonctionnement interprétatif qui leur est reconnu, au moins implicitement, par toutes les théories existantes. Qu'on parle plutôt d'unicité, de rôle, de désignation contingente, cela se laisse toujours ramener à la même chose. Pour un groupe nominal défini, il s'agit toujours de déterminer sur la base de son contenu descriptif un domaine d'interprétation et un critère de sélection tel que ce critère ne s'applique qu'à un élément du domaine.[1] »

À partir d'un rôle donné, *i.e.* à partir d'un type de domaine où elle est applicable[2], une description définie pourra avoir *n* valeurs (désigner *n* individus), autant de valeurs qu'elle peut construire de domaines d'interprétation. Autrement dit, si une description définie construit toujours la même classe, l'élément de la classe qu'elle désigne, lui, change d'un contexte à l'autre : « Ce qui est le propre des référents [syntagmes nominaux définis ici], ce n'est pas d'être dépourvus de sens constant, c'est que leur sens constant est une fonction. Ces signes ont bien une valeur sémantique stable et spécifique. Mais leur sens est de la forme générale *f(y)* où *y* est une variable parcourant l'ensemble des actes concrets d'énonciation.[3] »

C'est cette possibilité, pour une description définie, de pouvoir désigner des valeurs différentes qui la distingue crucialement du nom propre qui, lui, renvoie toujours à la même valeur. De ce point de vue, la description définie présente l'avantage, sur le nom propre, de pouvoir désigner un individu « inconnu ». Cette caractéristique est particulièrement précieuse pour la construction de l'objet de discours dans les chaînes de référence.

Que devient alors ce mécanisme caractéristique des descriptions définies lorsque les valeurs qu'elles désignent se trouvent être uniques ? Retrouve-t-on une mise en relation directe avec l'objet, typique du nom propre ? La notion de rôle, ou encore celle de classe, est-elle encore pertinente ?

B - Deux cas particuliers de descriptions définies

Il est des types de descriptions définies qui illustrent des cas de disjonction rôle/valeur : il s'agit d'une part (B1) des « descriptions définies complètes » et d'autre part (B2) de ce que Kripke nomme les « désignateurs rigides *de facto*[4] ».

B1 - Les descriptions définies complètes

Les descriptions définies complètes (ou encore descriptions identifiantes) se distinguent des descriptions définies incomplètes sur deux points :

- l'ordre rôle/valeur y est interchangeable. On reprend ici l'exemple fameux proposé par Donnellan [1971] : l'interprétation de la description définie *l'assassin de Smith* peut se faire soit en termes de valeur (on répond à la question « qui est l'assassin de Smith ? ») soit en termes de rôle (on répond à

[1] Corblin 1995, p. 192.
[2] Le type de domaine adéquat pour une description définie est en partie déterminé par la signification des éléments lexicaux qui la constituent, *supra*.
[3] Berrendonner 1978, p. 47.
[4] Pour une première approche, voir précédemment chapitre III § III.2.

la question « qu'a fait l'assassin de Smith ? », sans se préoccuper de l'identité du meurtrier). Dans ce cas, la description définie peut indifféremment pointer soit sur un objet soit sur une classe ;

- la définition proprement dite d'un domaine, c'est-à-dire la fonction de « filtre » tenu par le « rôle », ne semble pas y être nécessaire. En effet, il suffit qu'un seul paramètre soit fixé, indépendamment du contexte d'occurrence de la description définie, pour que la référenciation à l'objet soit réalisée. L'exemple le plus célèbre sur ce point est celui de la description définie *le président des États-Unis*, où seule la mention du « monde possible » où l'on se trouve, parce qu'elle fixe le paramètre temporel, est nécessaire à l'individuation.

C'est sur la base de ces deux propriétés, qui semblent permettre un lien direct avec le référent, sans qu'il y ait nécessité de former un domaine (une « classe »), que les descriptions identifiantes ont pu être rapprochées des noms propres. Or, il semble que les deux cas sont différents.

L'effet de référence directe est lié, dans le cas de la description identifiante, à l'existence unique de l'objet dans un monde donné. C'est pourquoi le domaine d'application de la description définie n'a pas besoin d'être spécifié : si l'objet existe dans ce monde possible, alors cet objet est celui que l'on connaît, quels que soient les énoncés où il peut être inséré. Alors que le domaine n'a pas besoin d'être précisé dans le cas des descriptions identifiantes, le domaine doit être oublié, sitôt le baptême prononcé, dans le cas des noms propres, ce qui est bien différent.

On peut en déduire que ce n'est pas l'unicité du référent dans un monde possible qui peut discriminer les descriptions définies (entre complètes et incomplètes). Dans les deux cas, il y a présence d'un domaine, qui est, d'un point de vue logique, soit actif (action en termes de « rôle » filtrant) soit inactif. Dans ce dernier cas, le domaine n'a pas besoin d'être précisé et la construction d'une classe peut n'y être que virtuelle.

Par l'examen du fonctionnement particulier des descriptions identifiantes, apparaît la possibilité, pour certains types de descriptions définies, d'avoir des emplois rigides[1]. C'est plutôt de ce point de vue qu'il nous paraît pertinent de distinguer les différents types de descriptions définies : si la plupart des descriptions identifiantes correspondent à des emplois rigides, il y a aussi des descriptions définies incomplètes qui peuvent connaître des emplois rigides. Nous y revenons en III.

B2 - Désignateurs rigides de facto

Dans le cas des désignateurs rigides *de facto*, les référents sont non seulement uniques dans un monde possible, mais ils le sont aussi dans tous les mondes possibles. Ainsi, dans le cas de la description définie « le plus petit nombre premier », il n'est pas même besoin, comme dans le cas des descriptions identifiantes, de fixer un paramètre pour individualiser l'objet. Pourtant, là encore, l'impression de relation directe avec le référent, sans passage par l'intermédiaire d'une classe, n'est que factice. En effet, comme nous l'avons précédemment relevé, et tout comme dans le cas des descriptions identifiantes, l'individuation du référent

[1] Voir, sur ce point, Fradin et Marandin 1979, p. 62 : « Un substantif réfère rigidement s'il porte un effet de référence au même objet dans n'importe quel discours où il est employé pour référer à un objet extra-linguistique ».

relève d'une connaissance extra-linguistique, et, d'une certaine façon, d'une contingence[1]. L'indifférence à la pluralité des mondes possibles y est donc le fruit d'un hasard et non, comme dans le cas du nom propre, d'une nécessité. Comme le fait remarquer Récanati [1983], la fonction directement individualisante du désignateur rigide *de facto* relève d'une connaissance qu'il n'est pas nécessaire d'avoir pour comprendre ou utiliser la description définie : « On peut très bien comprendre la phrase ["la racine carrée de 25 est un nombre impair"] sans savoir que "la racine carrée de 25" dénote 5 dans tous les mondes possibles, et donc sans savoir qu'elle est vraie si et seulement si 5 est impair.[2] »

À la suite de Fauconnier[3], on peut dire que les emplois rigides de certains types de descriptions définies (comme ici les désignateurs rigides *de facto*) viennent du fait que leur interprétation en tant que « rôle » se trouve être coïncidente avec leur interprétation en tant que « valeur ». Cette coïncidence ne doit pas, pour autant, laisser supposer que les descriptions définies identifiantes ou les désignateurs rigides *de facto* fonctionnent, d'un point de vue logique, sans passer par la construction d'une classe.

C - Précision sur la notion de classe

La coïncidence de l'interprétation rôle/valeur dans certains types de descriptions définies fait émerger la notion de classe virtuelle. Par classe virtuelle, on entend une classe dans laquelle un seul élément peut être effectivement désigné, les autres éléments de la classe construite étant virtuels, non identifiables. Le « contenu » d'une telle classe n'est pas fondamentalement différent de celui des classes construites par les autres types de descriptions définies. Si, d'un point de vue classiquement logique, une classe se laisse entendre comme un ensemble de propriétés générales communes à un ensemble d'objets, nous avons proposé, à la suite de Chastain et des reformulations linguistiques dont il a fait l'objet, de comprendre la « classe » d'un terme textuel en termes de contenu descriptif. Un contenu descriptif est alors constitué :

- de la « dénotation » au sens de Chastain, ou encore, dans une perspective linguistique[4], de la signification des éléments lexicaux constituant une description définie. C'est cette signification lexicale qui permet de sélectionner un domaine d'interprétation valide pour une description définie ;

- du complexe des relations qu'entretient une description définie dans une chaîne de référence avec d'autres objets du texte : c'est ce qui distingue la dénotation de la référence discursive, dans les termes de Chastain.

La notion de classe ainsi comprise ne suppose pas que la fonction singularisante des descriptions définies s'apparente à l'instanciation d'une variable : en ce sens, le nombre de variables (le nombre d'objets d'une classe) n'est pas ici pertinent pour distinguer entre les types de descriptions définies.

Avec la notion de classe conçue en termes de contenu descriptif, le référent discursif qui permettra, *in fine*, à un utilisateur d'effectuer un découpage référentiel

---

[1] Dans le sens où ce type de connaissance est éminemment variable d'un individu à l'autre.
[2] Récanati 1983, p. 117.
[3] Fauconnier 1984, p. 198.
[4] Voir, précédemment, la notion de référence virtuelle proposée par Milner.

(extra-linguistique) sera un individu qui aura d'autres propriétés que celles portées par la « dénotation » de l'unité linguistique qui lui a initialement permis l'accès à un texte : c'est en ce sens qu'il peut, nous semble-t-il, y avoir « information ». Ceci suppose que le descripteur puisse être capté au niveau de la « classe » qu'il permet de constituer, et pas uniquement au niveau de l'objet qu'il permet de désigner. C'est par une approche linguistique que nous tenterons de traiter cet aspect du descripteur.

II.3 - Conclusions intermédiaires

En nous inspirant du modèle des chaînes de référence proposé par Chastain et des commentaires linguistiques dont il a fait l'objet, nous avons proposé d'étudier le descripteur sous l'angle d'une unité susceptible d'appartenir à la fois à une chaîne anaphorique et à une chaîne référentielle.

À partir des propositions de Corblin, nous avons dégagé dans un premier temps deux candidats-descripteurs : le nom propre et les descriptions définies complètes, éléments constitutifs des chaînes référentielles.

A - Conclusions sur le candidat « nom propre »

D'emblée, le nom propre, ne pouvant appartenir qu'à une chaîne référentielle[1], semblait constituer un mauvais candidat-descripteur, du moins dans le cadre de l'indexation que nous défendons (indexation explicative). Néanmoins, nous devions étudier plus précisément le fonctionnement logique du nom propre notamment pour pouvoir expliquer sa présence massive dans les pratiques d'indexation courantes. À ce titre, nous avons fait émerger la notion de « nom propre pourvu de rôle » qui semble particulièrement à l'œuvre dans l'indexation interprétative : une étude plus précise reste cependant à mener.

Si, dans le cadre de l'approche du descripteur que nous proposons, le nom propre ne peut être retenu comme descripteur, il peut être possible de l'exploiter, notamment comme auxiliaire dans un système d'information. En effet, les noms propres constituent des « points fixes de référence » qui présentent l'avantage de ne nécessiter l'intervention d'aucune autre description pour que l'on puisse les utiliser. Pour peu qu'on les connaisse, les noms propres constituent en cela des contextes pertinents pour les descriptions définies : « Leur rigidité "désignationnelle" en fait des candidats idéals pour marquer les points ou indices référentiels que nécessite l'interprétation de toute description définie.[2] »

C'est ainsi que, dans ce que les logiciens nomment les descriptions identifiantes, on trouve le plus souvent un groupe prépositionnel constitué d'un nom propre : « le président des *USA* », « l'assassin de *Smith* », « l'ermite de *Croisset* », etc. La référence singulière est, dans ce cas, facilitée par les points référentiels rigides que constituent les noms propres.

Ce qui est ici valable dans le domaine d'interprétation de la description définie elle-même peut, peut-être, moyennant des spécifications, s'appliquer au domaine d'interprétation du « thème de discours ». À ce titre, disposer, pour un texte donné, de l'ensemble des noms propres de ce texte peut constituer un cadre spatio-

[1] Corblin 1995, p. 194.
[2] Kleiber 1981, p. 319.

temporel de nature à aider la construction de l'interprétation du thème discursif[1]. Reste la nécessité de maintenir distinctes les occurrences du nom propre où il est pourvu d'un rôle et celles où il n'a qu'une valeur : la référence à l'individu n'est pas, comme nous l'avons vu, de même nature.

B - Conclusions sur le candidat « description définie complète »

D'après la typologie proposée par Corblin, seules les descriptions définies complètes peuvent appartenir à la fois aux chaînes anaphoriques et aux chaînes référentielles. Compte tenu de la définition du descripteur que nous avons établie, seules les descriptions identifiantes seraient donc susceptibles de constituer de bons candidats-descripteurs.

Par l'examen du fonctionnement logique des descriptions définies complètes comme des désignateurs rigides *de facto*, nous avons tenté de montrer que ces deux types d'unités se distinguaient moins du point de vue de la singularité de leur objet que du point de vue de leur mode de référenciation. La notion de classe, même si une classe est dans ces cas virtuelle, caractérise le fonctionnement logique de ces unités, ce qui les rapproche des descriptions définies incomplètes et les distingue du nom propre. Ce qui, en revanche, peut les rapprocher du nom propre, c'est un type d'emploi que l'on a, à la suite de Récanati, proposé d'appeler « rigide ». Il nous semble que c'est essentiellement sous cet angle que l'on peut distinguer, parmi les descriptions définies, celles qui peuvent être utilisées comme descripteurs. Le rôle contextualisant du nom propre quand il apparaît dans le groupe prépositionnel d'une description identifiante constitue un premier indice : nous aurons à examiner, d'un point de vue linguistique, le fonctionnement des compléments dans les groupes nominaux.

III - Approche linguistique du descripteur

Dans sa phase de détermination de descripteurs, l'indexation consiste en une extraction d'unités de discours. Nous avons spécifié le rôle de ces unités extraites des textes en nous appuyant sur le modèle des chaînes de référence : ces unités doivent présenter la caractéristique de pouvoir être éléments à la fois de chaînes anaphoriques et de chaînes référentielles.

D'un point de vue logique, les unités d'indexation doivent être pourvues des propriétés suivantes :

(i) elles doivent être des termes singuliers, c'est-à-dire être dotées d'un pouvoir référentiel leur permettant d'identifier un objet, objet de discours et/ou objet mondain ;

(ii) elles doivent être, plus précisément, des descriptions définies, c'est-à-dire établir une relation indirecte, *via* une classe, avec l'objet qu'elles désignent.

---

[1] Par exemple, proposer la description « l'affrontement des forces armées » et le nom propre « Nigeria » peut être utile, mais encore faut-il que le rôle contextualisant du nom propre soit spécifié comme tel dans le système d'information.

Elles doivent donc pouvoir être une dénomination de classe, en plus d'être la signalisation d'un individu ;

(iii) elles doivent, en outre, en tant que descriptions définies, pouvoir connaître des emplois rigides, c'est-à-dire pouvoir être une dénomination de classe stabilisée. Sur ce point, nous faisons l'hypothèse qu'il n'y a pas que les descriptions définies complètes qui sont susceptibles d'établir, sur des bases communicatives, une désignation stable à une classe d'objets.

Il ne semble pas que la seule approche logique permette de spécifier la propriété (iii) du descripteur : d'un point de vue strictement logique, seules les descriptions identifiantes constituent de bons candidats-descripteurs. Nous faisons l'hypothèse que d'autres types de descriptions définies peuvent également fonctionner comme éléments de chaînes anaphoriques et de chaînes référentielles. C'est sur la base de critères linguistiques que l'on peut essayer de les déterminer et de spécifier ainsi la morphologie du descripteur dans le cadre d'une indexation de type explicatif.

Nous essayerons dans un premier temps de reformuler d'un point de vue linguistique la propriété logique (iii) que nous avons attribuée au descripteur (III.1). Nous explorerons ensuite deux modèles de représentation linguistique qui permettent de déterminer la morphologie du descripteur (III.2). Le modèle proposé par Michel Le Guern fournit le cadre général qui permet de définir le descripteur d'un point de vue logico-sémantique (III.2.1). Nous proposons ensuite de nous intéresser plus particulièrement aux propriétés syntaxiques qui doivent, selon nous, caractériser le descripteur : à ce titre, nous explorerons le modèle d'analyse de la synapsie proposé par Sophie David (III.2.2). Nous évoquerons en fin de chapitre les possibilités d'automatisation de telles procédures d'extraction d'unités de discours (III.3).

III.1 - La « rigidité » du descripteur

D'un point de vue logique, le descripteur se laisse appréhender sous la forme d'une description définie susceptible de connaître des emplois rigides. Précisons ce que, d'un point de vue linguistique, peut vouloir dire la notion d'emploi rigide d'une description définie.

A - L'identité d'interprétation

Rappelons tout d'abord que la rigidité dont doit pouvoir faire preuve le descripteur est liée au rôle que nous avons attribué à l'indexation : elle doit permettre d'établir une identité d'interprétation, opération indispensable à la construction du thème de discours. L'identité d'interprétation, telle qu'on peut la concevoir dans le cadre des chaînes de référence, peut être ainsi définie : « Il y a identité d'interprétation si *a* et *b* reçoivent la même interprétation en vertu de règles qui ne doivent rien à leur proximité dans le même segment linguistique.[1] »

L'identité d'interprétation se distingue de l'interprétation de reprise dans laquelle l'interprétation d'un élément *b* nécessite l'emprunt, à un terme proche *a*, d'un élément qui fixe son interprétation[2]. Autrement dit, l'identité d'interprétation est le propre des chaînes référentielles et l'interprétation de reprise le propre des chaînes

[1] Corblin 1995, p. 111.
[2] *Ibid.*, p. 112.

anaphoriques. Dans le premier cas, les unités en jeu pourront être dites « saturées » du point de vue interprétatif, tandis que, dans les chaînes anaphoriques, les formes seront dites incomplètes, c'est-à-dire non saturées : c'est la mise en relation au contexte qui les sature, en fixant, le plus souvent par emprunt, une dimension manquante de l'interprétation.

B - Rigidité de désignation à une classe

L'identité d'interprétation ne peut s'établir qu'entre unités autonomes du point de vue de leur contenu interprétatif : en cela, elles doivent être des unités « saturées », si l'on considère ici la saturation sous un angle non pas référentiel mais sémantique[1]. Corblin engage à distinguer la saturation de la tête nominale d'un groupe nominal du calcul référentiel qui engage le groupe nominal dans son entier ; de ce point de vue, on peut distinguer la rigidité de désignation à une classe et la rigidité de désignation à un individu de cette classe : la saturation sémantique correspond à la désignation rigide à une classe, la saturation référentielle à la désignation rigide à un élément d'une classe.

Quand la classe constitue un singleton (comme dans le cas des descriptions identifiantes et dans celui des désignateurs rigides *de facto*), la rigidité de désignation à la classe se confond avec la rigidité de désignation à l'individu : il s'agit là d'un cas particulier. Le descripteur est concerné par la désignation rigide à une classe et, à ce titre, c'est la saturation sémantique des descriptions définies qui nous occupera.

La saturation sémantique correspond au fait que le « domaine de référence », ou encore le « domaine d'interprétation », d'une description définie est complet, dans le sens où il ne requiert aucun emprunt à un autre élément.

C - Notion de saturation

La saturation d'un groupe nominal peut s'établir de plusieurs façons[2], que Corblin illustre par les trois exemples suivants[3] :
(8) Pierre *chassait. Le chien* partit au loin
(9) Pierre vit *un chien. Le chien* partit au loin
(10) *Le chien* de Pierre partit au loin

Dans les deux premiers exemples, les descriptions définies sont incomplètes, ou encore non saturées : leur contenu interprétatif « doit être saturé par association à un domaine dans lequel les parties non spécifiées de leur contenu sont tenues pour déjà fixées[4] ». C'est le cas typique des chaînes anaphoriques : dans (8), l'interprétation du groupe nominal « le chien » a besoin de recourir à l'élément « chassait », dans (9) à l'élément « un chien ». Dans (10), la description définie est complète, ou encore saturée : elle est autonome du point de vue interprétatif. En cela, elle peut apparaître dans les chaînes référentielles. Typiquement, le groupe nominal utilisé comme descripteur, c'est-à-dire comme élément de chaîne référentielle, doit contenir, dans le groupe nominal lui-même, tous les éléments

[1] Corblin 1995, p. 132 par exemple.
[2] *Ibid.*, p. 183 : « Saturer un groupe nominal serait donc, *sous des formes différentes, repérer des points de référence* qui permettent par relation, d'isoler des individus au moyen de la description ». C'est nous qui soulignons.
[3] *Id.*
[4] *Id.*

lexicaux lui permettant de disposer d'un domaine d'interprétation, d'un domaine de référence, complet, autonome par rapport à son contexte d'apparition. C'est ainsi que la forme linguistique privilégiée du descripteur est celle du groupe nominal « complexe[1] », pourvu de compléments, dont on montrera qu'ils ne sont pas nécessairement de la forme « de + Nom propre ». Par ailleurs, ce type de description définie peut tout à fait apparaître dans une chaîne anaphorique : elle peut servir par exemple à saturer une description définie incomplète[2].

Pour pouvoir être une dénomination de classe autonome, le descripteur doit donc être une description définie indépendante de la mention d'autres éléments : son contenu lexical doit suffire par lui-même à déterminer un domaine d'interprétation.

Reste que cette autonomie de contenu interprétatif, si elle est une condition nécessaire aux emplois rigides, n'en constitue pas une condition suffisante. D'autres critères entrent en jeu, notamment des critères de nature pragmatique, liés au fait que la relation référentielle de ce type de description définie s'établit sur des « bases communicatives », pour reprendre les termes de Corblin. Nous ne détaillerons pas ici ces autres critères : nous avons proposé de les traiter, sous un autre angle, au niveau de la constitution des documents[3].

Pour pouvoir extraire des descripteurs, nous devons donc disposer d'un modèle de représentation linguistique qui permette de capter une unité de discours de la forme « groupe nominal défini complexe ». Pour ne pas compliquer la présentation qui suit, on étudiera uniquement les descripteurs de la forme le N de N (type le centre de documentation), *alors que, bien entendu, d'autres formes de groupes nominaux complexes sont pertinentes en indexation : celles construites avec un adjectif, type le* NA (le traitement documentaire), *ou avec un nom, type le* NN (l'indexation matière), *etc. Notre étude du descripteur est ici conduite sous l'angle plus général de l'indexation : c'est la problématique du descripteur que nous abordons dans ce chapitre et non le détail de la variété morphologique qu'il peut prendre. Précisons enfin que si notre présentation privilégie, pour des raisons de clarté, une seule forme possible de descripteur, les modèles de représentation linguistique ci-dessous présentés proposent un traitement complet de tous les types de formes « complexes ».*

III.2 - Examen des modèles de description linguistique

Nous avons retenu deux modèles de représentation linguistique susceptibles de nous aider à spécifier la morphologie du descripteur dans le cadre d'une indexation de type explicatif :

- le premier, établi par Michel Le Guern, est le seul qui, à notre connaissance, propose une représentation linguistique du descripteur comme unité de discours : le descripteur y est explicitement défini comme un syntagme

---

[1] Nous entendons par là des groupes nominaux qui comportent des groupes prépositionnels comme dans « le traitement *de l'information* ». Nous revenons largement sur cet aspect ci-après.

[2] Par exemple dans le contexte suivant : « *Le chien de Pierre* partit au loin. On retrouva *la bête* écrasée ». La description définie incomplète « la bête » peut être saturée par recours à la description définie complète « le chien de Pierre ».

[3] C'est l'enjeu des propositions du chapitre IV où l'on tente de montrer que les bases communicatives par lesquelles vont pouvoir s'établir les chaînes référentielles doivent être déterminées en amont de l'extraction des descripteurs proprement dite.

nominal. La notion de syntagme nominal qui sous-tend ce modèle croise deux points de vue[1] : le point de vue syntaxique permet de faire voir le descripteur comme créé par/issu du discours lui-même ; le point de vue logico-sémantique permet de spécifier le comportement interprétatif du descripteur en discours : il établit une « relation référentielle *autonome*[2] » ;

- le second modèle, proposé par Sophie David, ne concerne pas le descripteur proprement dit, mais un type d'unité linguistique susceptible d'être utilisé comme tel (la synapsie). La représentation linguistique proposée ici diffère essentiellement de la première sur deux points : elle est menée dans le cadre strict de l'analyse linguistique, la référence à un domaine d'application (comme l'indexation) ne constitue pas le cœur du propos. La représentation qui y est proposée du groupe nominal est en outre menée du seul point de vue de la syntaxe. Cependant, dans le cadre d'analyse proposé, la syntaxe est comprise comme un mode d'organisation de l'interprétation[3] et rejoint en cela l'approche proposée par Michel Le Guern.

Comme nous essaierons de le montrer, les deux modèles de représentation se complètent :

- par le premier, on dispose d'un cadre théorique global, qui permet d'approcher la morphologie du descripteur sous deux points de vue, qui peuvent être tenus pour partiellement différents. Le point de vue de l'interprétation du descripteur en discours, qui correspond au point de vue de l'utilisateur ; le point de vue de l'extraction du descripteur, qui correspond au point de vue de l'indexeur (qui se situe en deçà de l'interprétation). Si le modèle de Michel Le Guern a principalement été exploré sous l'angle de la recherche documentaire, il permet, comme nous le montrerons, de spécifier aussi le descripteur sous l'angle de l'indexation elle-même. C'est sous cet angle que nous exploiterons le modèle de Michel Le Guern ; c'est sous cet angle aussi que certaines des descriptions proposées demandent à être précisées ;

- le second modèle vient compléter le premier en ce qu'il précise, nous semble-t-il, certains aspects de la morphologie du descripteur vue du côté de l'indexation. Par le biais de la description de la synapsie que propose Sophie David, on peut, nous semble-t-il, détailler la morphologie du descripteur, sans avoir à la redéfinir. Ce sont uniquement des critères syntaxiques qui seront ici proposés : leur pertinence documentaire, si elle a déjà fait l'objet d'investigation[4], nous semble pouvoir être théoriquement fondée.

En outre, l'intérêt de ces deux modèles est d'avoir fait l'objet d'implémentation informatique (III.3) : on dispose dans les deux cas de dispositifs automatisés d'extraction d'unités de discours. Les représentations linguistiques présentées ici peuvent donc être « utilisées » par les indexeurs dans une perspective professionnelle par le biais des systèmes informatiques dont elles ont fait l'objet. Sur ce point, signalons une différence entre les deux dispositifs automatisés, comme entre les deux modèles linguistiques qui les sous-tendent. Alors que le

---

[1] Pour Michel Le Guern, il importe de maintenir conjoints ces deux points de vue. Le Guern 1991a, p. 23 : « Que le syntagme nominal soit une structure syntaxique, c'est une évidence, mais il faut y avoir aussi une structure logico-sémantique ».
[2] Le Guern 1994, p. 75. C'est nous qui soulignons.
[3] Marandin 1992, Milner 1989.
[4] Par exemple à la Cour des Comptes dans le domaine de l'indexation, voir Simonot 1993.

dispositif issu du modèle de Michel Le Guern, établi au départ pour l'analyse du français, a fait l'objet de recherche pour une application dans d'autres langues (notamment l'arabe et le portugais), le système issu du modèle proposé par Sophie David ne traite que les textes écrits en français[1].

III.2.1 - MODÈLE LOGICO-SÉMANTIQUE PROPOSÉ PAR MICHEL LE GUERN

Comme nous l'avons indiqué à plusieurs reprises, cette recherche prend son ancrage dans le modèle du descripteur établi par Michel Le Guern, modèle que l'on a tâché d'appréhender sous un point de vue un peu différent, en mettant l'accent sur la référence discursive qui s'établit préalablement à la référence mondaine proprement dite, caractéristique de l'indexation. En cela, le rôle que Michel Le Guern attribue au descripteur, comme la représentation logico-sémantique qu'il en propose, nous paraissent tout à fait adéquats pour définir la morphologie générale du descripteur. Le modèle établi permet de mettre en valeur des formes linguistiques spécifiques : les syntagmes nominaux « complexes » constituent la forme linguistique privilégiée des descripteurs (A). Le modèle permet également de déterminer des niveaux d'appréhension différents du descripteur, selon le point de vue retenu (utilisateur/indexeur) (B). Sur ce point, l'approche de Michel Le Guern dégage un niveau de perception du descripteur qui pose de façon aiguë la problématique du descripteur comme unité de discours (C).

A - Le descripteur est un syntagme nominal, et, de façon privilégiée, un syntagme nominal « complexe »

Dans le cadre mixte, à la fois logique et grammatical, posé par Michel Le Guern[2], le descripteur est passible d'une double description :

- d'un point de vue logique, un descripteur est un « terme » : le terme logique s'oppose au « prédicat », comme un « objet » à ses « propriétés ». Un terme est un objet construit, résultat d'une opération de fermeture d'un prédicat par un quantificateur (voir ci-après). En tant que terme, il fonctionne comme une dénomination de classe, référentiellement autonome ;

- d'un point de vue linguistique, c'est un « syntagme nominal » et, plus particulièrement un syntagme nominal défini[3] ; il est à ce titre construit par la syntaxe. C'est en outre un syntagme nominal défini interprétativement « saturé » au sens proposé précédemment : en discours, les syntagmes nominaux les plus évidemment descripteurs sont ceux comportant des groupes prépositionnels.

  C'est en ce sens que la « forme normale » du descripteur est celle d'une unité composée de plusieurs mots[4] (type « le traitement de l'information »), le descripteur simple apparaissant dans ce cadre comme une exception[1].

---

[1] Rappelons que, dans cette recherche, nous avons uniquement traité l'indexation des textes écrits en français.
[2] Le Guern 1984, 1989, 1991a, 1991b, 1994.
[3] En effet, il semble que le descripteur soit, dans le cadre proposé par Le Guern, proche des « descriptions définies » de la logique, c'est-à-dire qu'il fonctionne de façon privilégiée comme s'il était précédé d'un article défini.
[4] Mais toutes les unités composées de plusieurs mots ne sont pas construites, dans ce modèle, par la syntaxe, voir ci-après la distinction entre les SN construits avec SP et les SN construits

Les deux dénominations rendent compte, sous une forme différente, de la même propriété définitoire du descripteur : c'est un « mot de discours », c'est-à-dire un mot construit par le discours qui révèle la même propriété que celle qui a été précédemment attribuée au « terme textuel ».

En tant que « légisigne[2] », il n'est pas seulement une occurrence[3], il est aussi un « type », et un « type » issu non de la langue mais du discours. En effet, en tant que « légisigne indiciaire » et non « symbolique », le descripteur comme « type » n'est pas l'interprétant de ses occurrences, il en est le « nom propre[4] », dans le sens où l'on a vu qu'une description définie pouvait connaître des emplois rigides (référence stabilisée à une classe d'objets). L'emploi rigide du syntagme nominal qui lui permet d'être interprété comme dénomination de classe stable correspond à l'emploi documentaire du syntagme nominal, c'est-à-dire au descripteur.

L'approche logico-sémantique du syntagme nominal, dans laquelle Michel Le Guern définit le descripteur, fait apparaître une particularité totalement ignorée des définitions classiques du descripteur : la morphologie canonique du descripteur est celle du syntagme nominal « complexe », en vertu des propriétés interprétatives constitutives du syntagme nominal lui-même.

Cette approche, qui fait apparaître la morphologie caractéristique du descripteur comme unité de discours, privilégie le point de vue de la recherche documentaire[5] et non celui, à proprement parler, de l'indexation : à quel niveau l'indexation doit-elle capter le descripteur ? Si, un utilisateur, construisant un objet de discours, ne peut, comme nous l'avons vu, que recourir à une forme de descripteur qui est celle du groupe nominal complexe, il faut bien voir que l'indexeur, lui, doit proposer, à l'interprétation, le descripteur sous une forme quelque peu différente. Le modèle de Michel Le Guern permet de spécifier la morphologie du descripteur du point de vue particulier de l'indexation que nous avons retenue (indexation explicative).

---

avec EP. Par contre, dans tous les cas, ce sont des descripteurs, au sens que Michel Le Guern donne à ce mot.

[1] On veut dire par là qu'entre une forme « simple » et une forme « complexe » (composée de plusieurs « mots »), on retiendra de façon privilégiée la forme composée (par exemple, « le traitement de l'information » et non « le traitement » et « l'information »). Dans ce modèle, on n'exclut pas de retenir comme descripteurs des formes simples quand elles se présentent comme telles dans un syntagme nominal, par exemple *l'indexation* dans : « Cette étude porte sur l'indexation ». Nous reviendrons sur cet aspect du descripteur comme « forme simple ».

[2] Voir précédemment chapitre II § III.2.

[3] Le Guern 1991a, p. 23 : « On pourrait le voir [le descripteur] comme le mot de la langue actualisé dans le discours mais cela n'est pas tout à fait suffisant ».

[4] Voir la notion de nom propre chez Peirce, évoquée au chapitre II § III.2.

[5] Le Guern 1994, p. 75 : « L'analyseur morpho-syntaxique élaboré par l'équipe SYDO a eu comme premier objectif l'extraction de tous les syntagmes nominaux présents dans le texte à indexer, ces syntagmes nominaux étant appelés à *jouer le rôle des descripteurs dans le système d'information* ». C'est nous qui soulignons.

B - Le descripteur correspond à un prédicat lié, et, plus spécifiquement, à un prédicat lié ouvert

Pour qu'un descripteur puisse désigner à la fois une classe et un objet, être à la fois « syntagme-type » et « syntagme-occurrence », il devrait, semble-t-il, relever d'un « niveau intermédiaire » que Le Guern note N' dans le cadre du modèle suivant[1] :

- le niveau N correspond au niveau de la langue et relève d'une logique intensionnelle : s'y trouvent les mots de la langue désignant des propriétés ;
- le niveau N'' correspond au niveau des objets et relève d'une logique extensionnelle : s'y trouvent les mots du discours désignant des objets ;
- le niveau N' correspond au niveau des « classes » et relève d'une logique extensionnelle : s'y trouvent des mots du discours désignant des objets et des classes, de façon virtuelle ou actuelle[2].

L'articulation de ces trois niveaux se laisse figurer de la façon suivante :

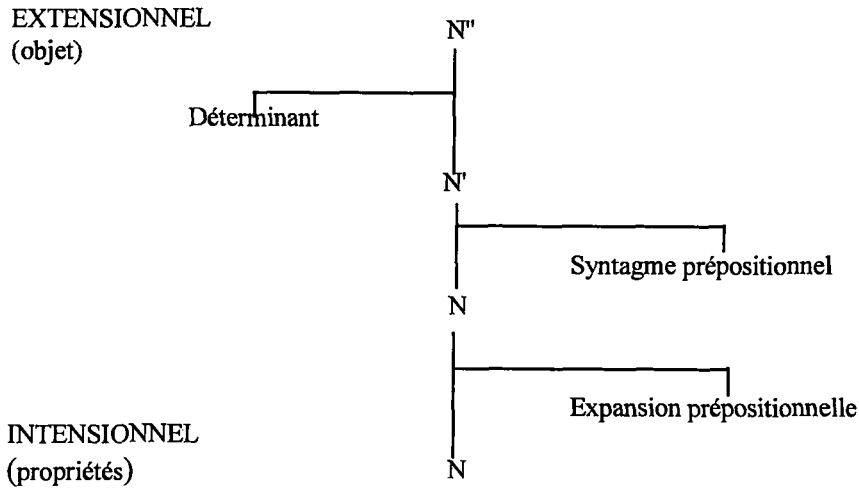


Figure 5 - Grammaire du syntagme nominal
Bouché 1989, p. 432

[1] Si Michel Le Guern propose de placer le descripteur au niveau N'', c'est essentiellement pour des raisons de clarté. Du point de vue théorique de l'indexation, le niveau N' paraît, comme le confirme Michel Le Guern, plus adéquat ; il ajoute que, d'un point de vue strictement documentaire et pratique, le niveau N'' n'est pas faux et semble plus communément admissible (communication personnelle).

[2] Le Guern 1991a : « Le passage du niveau N au niveau N' correspond à la prise en compte d'un univers donné, au surgissement de la référence, à la possibilité de déterminer des classes, au moins virtuelles ; c'est le basculement de la logique intensionnelle à la logique extensionnelle, c'est la mise en relation entre les mots et les choses ».

À chacun de ces niveaux correspond un type d'individu logique, mais pas nécessairement une forme linguistique spécifique :

- au niveau N, on trouve des « prédicats libres », libres de tout lien référentiel ; exemples[1] : *maison, soleil,* mais aussi *pomme de terre.*
- au niveau N", on trouve des « termes », c'est-à-dire des prédicats liés fermés par l'application d'un opérateur (un déterminant) ;
  exemples : *la* maison de Jean, *le* soleil, *les* pommes de terre
- au niveau N', on trouve des prédicats liés ouverts, c'est-à-dire des prédicats liés référentiellement mais non discriminants (pas de pointage sur un objet singulier) ;
  exemples : maison *de Jean*, soleil, pomme de terre.

Le niveau N' correspond, nous semble-t-il, au niveau où le descripteur prend toute sa pertinence documentaire. Le prédicat lié ouvert, associé à un univers discursif particulier qui spécifie son extension, peut alors fonctionner comme élément de chaîne référentielle, par laquelle de nouvelles propriétés (issues des discours) vont pouvoir venir alimenter les propriétés constitutives de sa classe. C'est, nous semble-t-il, ce phénomène que Michel Le Guern exprime en ces termes : « Alors que les sèmes sont pertinents pour le lexique, les emplois en prédicats liés peuvent se charger, par un effet de contexte, des traits de substance des classes d'objets qu'ils désignent, même s'ils n'ont pas été retenus par la langue pour constituer les signifiants lexicaux.[2] »

Le prédicat lié ouvert peut fonctionner ainsi dans la mesure où il réfère de façon autonome à une classe d'objets, c'est-à-dire dans la mesure où il fait « intervenir un objet de l'univers de référence ». Comme nous l'avons vu, l'intervention de cet l'univers de référence peut se faire au sein même du syntagme nominal par le biais des groupes prépositionnels : « traitement de *l'information* », « indexation *des documents* ». Cependant, comme dans le cas du nom propre utilisé classiquement en position « complément » dans les descriptions identifiantes des logiciens, la description définie comprise dans le groupe prépositionnel n'est pas tout à fait équivalente à la description définie employée seule (« l'information », « les documents »). Nous y reviendrons : cet aspect mérite en effet une attention particulière notamment pour montrer que l'indexation par « traitement » et « information » n'est pas équivalente à l'indexation par « traitement de l'information ».

Le modèle logico-sémantique établi par Michel Le Guern permet de pouvoir circonscrire un lieu précis dans lequel le descripteur fonctionne dans ce que nous avons proposé d'appeler, à la suite de Chastain et de Corblin, les chaînes référentielles : en tant que prédicat lié ouvert, unité du niveau N', le descripteur apparaît comme une dénomination de classe, susceptible d'être stabilisée dans la mesure où elle est référentiellement autonome. L'utilisateur exploitera cette fonction du descripteur pour « parcourir » des textes différents ; il considérera le descripteur sous la forme d'une unité N" lors de la lecture d'un document proprement dite : il sera alors nécessaire pour lui d'identifier non plus une classe mais un élément de cette classe.

Examinons de plus près les unités qui apparaissent au niveau N' : peuvent-elle toutes, au même titre, constituer des descripteurs au sens que nous avons défini ?

---

[1] Exemples repris de Le Guern 1991a.
[2] Le Guern 1989, p. 342.

C - Examen des unités captées au niveau N'

Il y a, dans le modèle proposé par Michel Le Guern, deux façons d'accéder au niveau N', étant posé qu'on ne peut directement passer du niveau N au niveau N" dans le cadre proposé[1] :

- soit N' –> N
- soit N' –> N + SP, avec SP –> P' + N"

S'il y a deux façons de construire N', il n'y paraît rien au niveau de N", qui se construit sur la base d'une seule et même règle : N" –> D' + N', où D' est le déterminant. Autrement dit, au niveau N", les descripteurs pourront être aussi bien de forme simple que de forme complexe, la forme complexe pouvant être soit du type « l'armée de *l'*air » (avec déterminant), soit du type « l'armée de terre » (sans déterminant), la forme simple pouvant être « l'air », par le jeu des règles de réécriture. Ces trois types de forme, quand elles sont captées au niveau N', sont-elles dotées des mêmes propriétés ?

Nous examinerons d'abord le cas des descripteurs « complexes » qui apparaissent au niveau N', puis le cas des descripteurs « simples » captés au niveau N'.

C1 - Les descripteurs « complexes » du niveau N'

Au niveau N', on peut avoir des descripteurs complexes de deux types, issus des deux règles ci-dessus mentionnées :

(i) le descripteur « armée de terre » est issu de la règle N' –> N, où N est de la forme « N + EP », avec EP –> P' + N. Dans « armée de terre », « de terre » se lit comme un EP, non construit par la syntaxe ;

(ii) le descripteur « armée de l'air » est issu de la règle N' –> N + SP, avec SP de la forme « P' + N" », où N" –> D' + N'. Dans « armée de l'air », « de l'air » se lit comme un SP, construit par la syntaxe.

La différence entre SP et EP tient à la présence ou à l'absence de D' qui, d'un point de vue linguistique, correspond à un déterminant et qui, d'un point de vue logique, comprend un « présupposé d'existence ». Sans le déterminant indiquant un présupposé d'existence, un N ne peut construire un objet de discours.

Aussi le descripteur (ii) se distingue-t-il du descripteur (i) en ce qu'il englobe une unité de niveau N" (ici « l'air »). Autrement dit, que l'indexation se place au niveau N' ou N", on devrait dégager du descripteur (ii) un autre descripteur, soit « l'air » si l'on se place au niveau N", soit « air » si l'on se place au niveau N'. Une indexation qui voudrait ne se situer qu'au niveau N' rencontre donc deux ordres de difficulté : doit-elle considérer que le descripteur complexe « armée de l'air » constitue un bon descripteur ? Devra-t-elle lui préférer le descripteur « air » par exemple ?

Ces deux questions se ramène à celle du « statut objectal », comme le formule Michel le Guern, du second N" englobé dans un N'. Dans EP, le N dépourvu de déterminant n'a clairement pas de statut objectal ; dans SP, où le N' est doté d'un déterminant, quel est le statut objectal de ce N" ? Le syntagme nominal englobé

---

[1] Où SP se lit « syntagme prépositionnel » et correspond à « de Jean » dans l'exemple « la maison de Jean » ; où la notation P' correspond à un élément qui peut être soit une locution prépositionnelle, soit une préposition (la catégorie morphologique de la préposition est notée par P), auquel cas on applique la règle P' -> P.

dans un autre syntagme nominal semble avoir un statut particulier : « Dans le cas du prédicat lié qui comprend un syntagme prépositionnel, et donc un syntagme nominal à l'intérieur de ce syntagme prépositionnel, on a bien d'une certaine manière un second objet, mais ce second objet perd son statut d'objet dès lors qu'il entre dans la composition d'un prédicat, ce qu'expriment clairement les grammaires d'opérateurs.[1] »

Il n'est donc pas évident, comme le dit Michel Le Guern, que ce second N" présente une « pertinence documentaire[2] ». Sur ce point, il n'est pas sûr que l'on doive privilégier, comme descripteur du niveau N', l'unité « air » au détriment de l'unité « armée de l'air ». Il faut tenir compte du fait que tous les syntagmes nominaux inclus dans un SP ne se laissent pas interpréter de la même façon.

Certains construisent un objet de discours, d'autres ne permettent pas véritablement d'en construire un ; tout dépend du rôle logique tenu par le défini inclus dans le syntagme prépositionnel. Soit le défini fonctionne comme un quantificateur et il construit un objet de discours, soit le défini est privé de sa fonction de quantificateur et il ne peut créer un objet de discours[3].

Dans les cas où le second N" construit un objet de discours, le syntagme nominal qui l'englobe n'aura pas forcément le statut de dénomination de classe : en effet, le second N" « sature » moins le syntagme nominal qui le contient, dans la mesure où lui-même nécessite un emprunt à d'autres éléments pour être saturé (on rejoint ici le cas des descriptions définies incomplètes, comme « le placard de la cuisine » ou « le chat du voisin »). Quand le second N" perd son statut objectal, le syntagme nominal englobant aura, lui, plus nettement le statut de dénomination de classe : c'est le cas typique des descriptions identifiantes (« le président des USA »), mais c'est aussi le cas des autres descriptions définies comme « la rose des vents[4] ».

On peut considérer que la perte du statut objectal du second N" (« les vents » par exemple) est lié à l'usage, on enregistre alors la forme dans un dictionnaire où il sera traité comme un N : c'est l'option retenue, sur un plan pragmatique, par les membres de l'équipe SYDO. On peut considérer aussi que le second N" n'a plus de statut objectal sur la base non seulement de critère d'usage, mais aussi de critère syntaxique : on se place alors dans une perspective qui cherche à expliquer le caractère lexicalisé ou lexicalisable des unités « complexes ». C'est l'option que retient Sophie David et que nous aborderons ci-après.

Dans notre approche du descripteur capté au niveau N', il nous semble en effet important de pouvoir distinguer, en toute généralité, parmi les syntagmes nominaux construits avec un SP, ceux où le second N" construit un objet de discours et ceux où il n'en construit pas. Ce point, crucial pour la détection des descripteurs conçus comme dénominations de classe, reste particulièrement délicat à traiter, notamment

---

[1] Le Guern 1991b, p. 80, n. 10.
[2] Ou, du moins, ce second N" présente une pertinence documentaire plus faible que le N" englobant (communication personnelle).
[3] Le Guern 1991a, p. 29 : « La vraie question, à propos des articles, est plutôt celle de l'article vide, que son insertion dans les formules figées prive de sa fonction de quantificateur : "rose *des* vents" n'implique plus qu'il y ait des vents ; la lexicalisation de l'expression fait que l'article *les* (des = de + les) ne porte plus ici le présupposé d'existence ».
[4] Exemple repris de Le Guern 1991a. *Id.*

du fait que, comme Berrendonner [1995a] a pu le montrer[1], certains syntagmes nominaux construits avec un SP peuvent être passibles de deux lectures, l'une en termes d'« objet », l'autre en termes de « type » : « [le N] peut renvoyer aussi bien à un objet extensionnel pris dans R [pour univers extensionnel] *(la porte du bureau, les archives du cinéaste Abel Gance...)* qu'à un type élément de I [pour univers intensionnel] (*le ministère de la guerre, le salon de l'automobile, la promotion de la femme...*).[2] »

On considérera que, du point de vue de l'indexation réalisée au niveau N', seront descripteurs les unités complexes qui construisent des « types », c'est-à-dire, dans le modèle de Michel Le Guern : soit des unités N' comportant un EP[3], soit des unités N' comportant un SP dont le second N" ne construit pas d'objet de discours. C'est la condition de « saturation » attendue du descripteur, compris comme élément d'une chaîne de référence, qui nous conduit à restreindre le type d'unités de discours susceptibles d'être descripteurs.

Comme Michel Le Guern, nous privilégions les syntagmes nominaux englobants, mais, contrairement à lui, dans une perspective qui est plus celle de l'indexation que celle de la recherche d'informations, nous proposons de distinguer, parmi les syntagmes nominaux englobants, ceux qui sont interprétativement saturés de ceux qui ne le sont pas.

Pour pouvoir établir cette distinction entre types d'unités construites avec un SP, nous observerons la description linguistique que Sophie David propose de la synapsie, § III.2.2 ci-après.

C2 - Les descripteurs « simples » captés au niveau N'

Les descripteurs de forme simple (constitués d'un seul « mot ») captés au niveau N' sont, eux aussi, issus des deux types de règle précédemment mentionnés. On peut donc avoir, au niveau N', une même forme de descripteur simple, issue de différents types de syntagmes nominaux définis.

Ainsi, par exemple, on obtient au niveau N' le même descripteur « indexation », qui peut provenir soit de (11) soit de (12):
(11) Cette étude porte sur l'indexation.
(12) Cette étude aborde l'indexation assistée par ordinateur.

En (11), « indexation » provient d'une description définie incomplète : le domaine de référence que construit l'unité au niveau N' s'établit sur la base d'un emprunt à d'autres éléments du contexte. On peut dire que la classe créée ici n'est que virtuellement construite, dans le sens où elle ne tient que par une interprétation de reprise, dans les termes de Corblin. Pour les raisons précédemment évoquées, une

---

[1] Comme nous l'a fait remarquer Michel Le Guern, Berrendonner se place ici dans la perspective d'une linguistique de la production plus que dans celle d'une linguistique de la réception (communication personnelle).
[2] Berrendonner 1995a, p. 20.
[3] Comme l'indique Berrendonner [*Id.*], les syntagmes nominaux construits avec EP pointent exclusivement sur des « types ».

telle unité ne peut être retenue comme descripteur, dans le sens que nous avons défini[1].

Le cas est différent en (12) où l'unité « indexation » provient d'une description définie qui peut être considérée comme complète, c'est-à-dire saturée du point de vue de son contenu interprétatif: les éléments lexicaux en position « compléments » ne requièrent pas d'emprunt contextuel, ce qui permet à l'ensemble de l'unité de pouvoir fonctionner comme dénomination de classe. Extraire dans ce cas l'unité « indexation » peut être intéressant, puisqu'en discours elle correspond à un référent discursif pour lequel un domaine d'interprétation est spécifié. Par ailleurs, le point d'accès que pourrait constituer une telle unité permettrait également à un utilisateur d'accéder à des expressions du type: « système d'*indexation assistée par ordinateur* » qui constitue une ressaisie du référent discursif précédent susceptible de permettre la construction de l'information.

Si, du strict point de vue de l'indexation, les descripteurs de forme simple ne constituent pas de bons candidats descripteurs, un système de recherche documentaire devrait sans doute pouvoir exploiter, d'une manière qu'il importe de spécifier précisément, un certain type de descripteurs de forme simple captés au niveau N': ceux qui constituent la « tête » d'une umité N' de forme « complexe ». Il faut donc pouvoir montrer que l'unité « indexation » dans « indexation assistée par ordinateur » fonctionne de la même façon que l'unité « indexation » dans « système d'indexation assistée par ordinateur ». La parenté des deux structures se laisse voir à travers la représentation linguistique de la synapsie que propose Sophie David.

En explorant le modèle logico-sémantique proposé par Michel Le Guern sous l'angle de la seule indexation, on peut définir la morphologie du descripteur sous ses principaux aspects. C'est un syntagme nominal complexe, qui, capté au niveau N', correspond à deux types de séquence :

- *les séquences « N + EP », qui constituent toutes des descripteurs ;*
- *les séquences « N + SP », dont certaines constituent des descripteurs : celles dont le second N" ne construit pas d'objet de discours.*

Nous allons ci-après examiner le modèle proposé par Sophie David sous l'angle de ces deux types de séquences, notamment dans le but de disposer de critères qui nous permettent de distinguer entre les séquences pourvues de SP.

Précisons d'ores et déjà que le modèle général du syntagme nominal[2] adopté par Sophie David étant différent de celui retenu par Michel Le Guern, les deux types de séquences dégagés ici ne se laisseront pas représenter exactement sous la même forme.

---

[1] Reste que, dans une perspective qui considère prioritairement la recherche d'information, il peut être important d'établir des moyens pour permettre à un utilisateur de saturer de telles unités (voir, sur ce point, le travail de Vidalenc-Sabourin 1989). Dans le cadre de notre approche qui cherche à considérer uniquement les faits d'indexation, de telles unités ne peuvent constituer de bons candidats-descripteurs.

[2] Qui reçoit plus précisément la dénomination de « groupe nominal » dans ce modèle. Mis à part les différences d'options théoriques qu'elles révèlent, les deux dénominations « syntagme nominal » et « groupe nominal » peuvent être tenues pour identiques.

III.2.2 - MODÈLE SYNTAXIQUE PROPOSÉ PAR SOPHIE DAVID

Nous présenterons, dans un premier temps, la représentation linguistique de la synapsie que propose Sophie David (A) ; nous discuterons, dans un second temps, la pertinence de cette représentation dans le cadre de notre approche du descripteur (B).

A - Représentation linguistique de la synapsie

La thèse de Sophie David[1] est consacrée à la description des unités nominales polylexicales[2] susceptibles de se lexicaliser, c'est-à-dire susceptibles de fonctionner comme une dénomination de classe[3]. Cette description, qui s'attache principalement aux unités construites par la syntaxe, rend compte du fonctionnement d'un type d'unités particulier : les « synapsies[4] ».

Le fonctionnement des synapsies sera plus compréhensible si le cadre syntaxique qui permet de les faire « voir » est rapidement et succinctement exposé.

Le cadre retenu dans cette analyse est celui de la syntaxe positionnelle, proposé par Milner[5] et repris par Marandin[6]. La syntaxe y est définie comme un « module », propre à construire des structures abstraites sur la base de « positions[7] ». La position est en effet l'unité de base de la syntaxe : dotée d'un certain nombre de propriétés[8], elle est rendue visible essentiellement par son occupation[9] par un « terme[10] ».

---

[1] Principalement David 1993a, mais aussi David et Plante 1990a, 1990b, 1991, et David et Souchard 1995.

[2] Une unité nominale polylexicale est une unité de catégorie nominale constituée de plusieurs mots et fonctionnant comme un mot simple, exemple : *pomme de terre* ; voir David 1993a, p. 8.

[3] « Une dénomination stabilisée régulièrement employée par une communauté de locuteurs » ; pour une description plus complète de la lexicalisation, voir David 1993a et Corbin 1992.

[4] La terminologie vient de Benveniste 1974 [1966], p. 163-176. Sophie David se distingue de Benveniste sur les propriétés qu'elle attribue à la synapsie, David 1993a, p. 35-37 et p. 130-132.

[5] Milner 1989.

[6] Marandin 1992b.

[7] Une position n'est pas une « place ». La première est de nature géométrique (une position est un ensemble de points géométriquement structurés), la seconde relève d'un ordre linéaire (David 1993a, p. 61). Milner (1989, p. 298) illustre la différence par l'opposition suivante : dans *j'ai fait prendre le train à mon fils* et *j'ai fait donner une couchette à mon fils*, *à mon fils* se trouve à la même place mais n'occupe pas la même position. Nous ne pouvons détailler ici la notion linguistique de position ni celle de syntaxe positionnelle, sur ce point voir Milner 1989.

[8] Ce sont des propriétés positionnelles et non lexicales, certaines d'entre elles étant interprétatives, voir sur ce point Milner 1989, p. 357 et suiv.

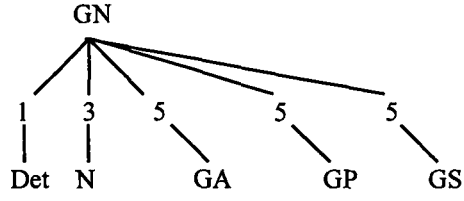
[9] Mais, ce faisant, la syntaxe, en tant que configuration de positions, n'est, elle, plus visible, et c'est là le paradoxe de la syntaxe, Kerleroux 1996, p. 389 : « On pourrait redire, selon les termes de Milner, que la syntaxe est essentiellement invisible : en effet, les positions, en tant que telles, sont non perceptibles par les sens, un espace géométrique ne rend ses cases visibles que lorsqu'elles sont occupées par des positions tangibles. Mais lorsqu'elles sont occupées, elles se confondent pour nos sens avec les termes occupateurs eux-mêmes ».

[10] La notion de terme est, chez Milner, globalement équivalente à celle d'unité lexicale, qui est définie par trois traits : une unité lexicale est une unité qui a une forme phonologique, une catégorie et une signification, Milner 1989 p. 324 et suiv.

Dans le cadre positionnel, un groupe nominal (GN) est une organisation qui a trois positions :

- une position noyau, nécessairement occupée par un constituant qui joue le rôle de « tête » dans le GN[1] ;
- une position de « spécifieur » à gauche de la tête, occupable par un déterminant ;
- une position à droite de la tête, occupable par des catégories majeures de type GA (groupe adjectival), GP (groupe prépositionnel), GS (relative) et GD (groupe adverbial).

La représentation positionnelle du GN « le fauteuil rouge de ma sœur qui a été réparé » sera de la forme[2] :

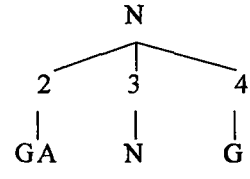


Où : « le » occupe la position étiquetée *Det*, « fauteuil » la position étiquetée *N*, « rouge » la position étiquetée *GA*, « de ma sœur » la position étiquetée *GP* et « qui a été réparé » la position étiquetée *GS*.

Dans ce cadre, la synapsie est définie par la conjonction des trois propriétés suivantes :

- c'est une unité de catégorie N ;
- c'est une unité qui occupe la position noyau du GN ;
- c'est une unité dont l'organisation syntaxique est, elle-même, constituée de trois positions : une position centrale qu'occupe le terme nominal jouant le rôle de tête, une position à gauche de la tête pouvant accueillir des constituants de type GA et une position à droite de la tête occupable par des catégories de type GA, GP et GS.

La représentation positionnelle de la synapsie « haute cour de justice » aura donc la forme :



Où : « haute » occupe la position étiquetée *GA*, « cour » la position étiquetée *N* et « de justice » la position étiquetée *GP*.

---

[1] Pour cette présentation, on se limitera au cas où cette position est occupée par un terme de catégorie N (nom).Cette position peut aussi être occupée par un A (adjectif) ou un V (verbe).
[2] Repris, comme l'ensemble des informations de ce paragraphe, de David 1993a, ici p. 161 ; la représentation adoptée emprunte le formalisme des arbres polychromes, présenté dans Cori et Marandin 1993 (cf. étiquetage chiffré des branches dans l'arbre des exemples), que nous ne développerons pas ici.

Cette représentation positionnelle du GN et de la synapsie (voir le schéma ci-dessous) montre que :

- la position d'un noyau GN peut être occupée soit par une unité simple (c'est le cas de « fauteuil » dans le GN « le fauteuil rouge de ma sœur qui a été réparé ») soit par une unité « complexe », une synapsie (comme « haute cour de justice », que l'on pourrait intégrer dans un GN de type « La *haute cour de justice* qui s'est prononcée hier »).

- les compléments à droite de la tête nominale sont différenciés : par exemple « de ma sœur » et « de justice » ne sont pas analysés de façon identique. Seul le GP « de justice » construit ici une synapsie, une unité pouvant servir de dénomination stable.

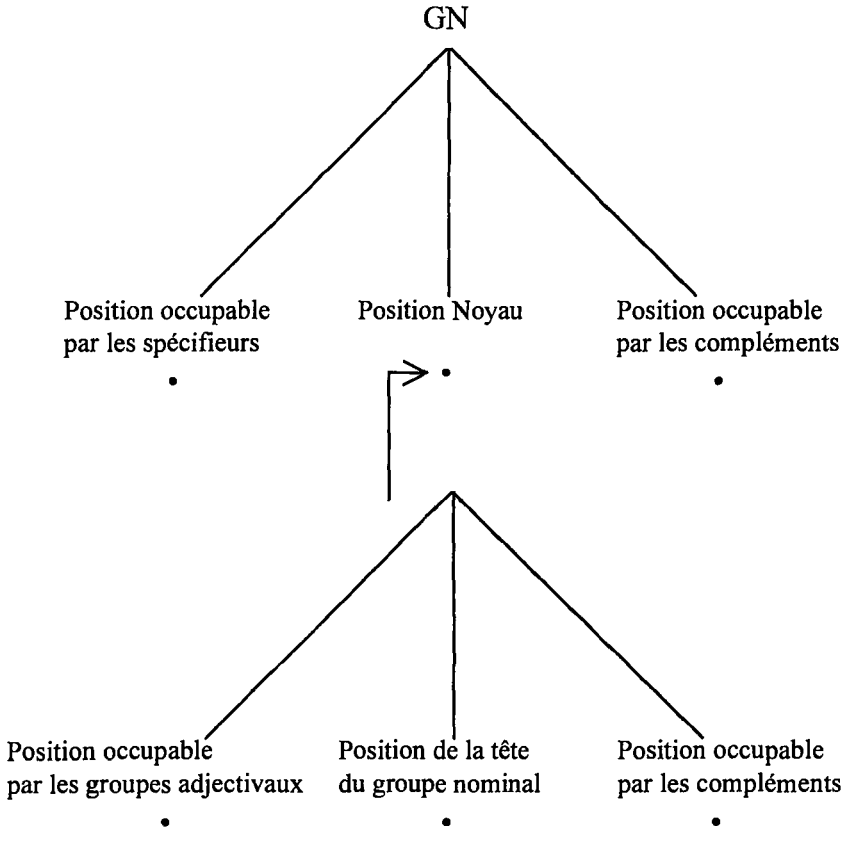


Figure 6 - Organisation positionnelle du groupe nominal et de la synapsie
David 1993a

Le marquage de la différence de complémentation à droite de la tête nominale, indiquée ici en termes d'unité (« GN » ou « synapsie »), recoupe, en partie, la différence notée, dans le modèle logico-sémantique de Le Guern, par la distinction EP/SP, relative au groupe prépositionnel.

Si l'observation et l'interprétation du phénomène sont relativement proches, en revanche, les critères formels qui permettent d'effectuer son repérage sont, eux, de nature différente.

Dans le cadre des deux modèles, c'est l'interprétation référentielle ou non du GP situé à droite de la tête nominale qui engage à établir une distinction, soit en termes d'unités, soit en termes de nature des compléments. Ainsi, dans le cadre positionnel, dira-t-on que :

- les compléments de la synapsie sont toujours aréférentiels[1] : ils ne permettent pas de construire un objet de discours ;
- les compléments du groupe nominal sont eux toujours référentiels : ils peuvent construire un objet de discours.

Cette différence peut être rendue visible par le biais d'un test, celui de la reprise anaphorique. Si une anaphore est possible, alors il est possible de construire un objet de discours :

- on ne peut avoir de reprise anaphorique dans le cas des compléments de la synapsie. L'unité « textes » dans « traitement de textes » ne peut construire un objet de discours : « *J'ai acheté un nouveau traitement de textes. Ils sont mieux écrits » ;
- on peut avoir une reprise anaphorique avec les compléments du GN : « ma sœur » issu du GN « le fauteuil rouge de ma sœur qui a été réparé » peut faire l'objet d'une anaphore et construire un objet de discours : « J'ai vu le fauteuil rouge de ma sœur qui a été réparé. Elle est contente du travail qui a été fait ».

Pour Sophie David, c'est précisément le caractère aréférentiel des compléments qui permet la constitution d'une unité particulière, qu'elle nomme synapsie, ayant la propriété spécifique de pouvoir fonctionner comme une dénomination de classe : c'est en effet parce que « blanc » est aréférentiel dans « ours blanc » que l'expression peut désigner une espèce d'ours sans que tous les ours qui en fassent partie ne soient nécessairement de couleur blanche.

Le problème est, comme nous l'avons précédemment évoqué, que le caractère aréférentiel des compléments de la synapsie n'est pas toujours inscrit de façon très nette dans le matériel linguistique perceptible.

Notamment, le critère du « déterminant » ne semble pas toujours suffisamment discriminant, surtout à l'intérieur d'un groupe prépositionnel. Si, en effet, l'absence de déterminant est un indice favorable à l'identification d'une synapsie (par exemple, « traitement *de* textes »), la présence d'un déterminant ne peut, à elle seule, indiquer, *a contrario*, que le complément est référentiel[2] et donc inapte à construire une synapsie (exemple, « robe du soir[3] »).

---

[1] Plus précisément et en reprenant les termes de Milner, on dira que les synapsies sont dotées d'une référence virtuelle mais pas de référence actuelle, alors que le GN a, lui, nécessairement une référence actuelle.

[2] David 1993a, p. 143 : « Ce n'est pas tant la présence du déterminant qui est importante, que la valeur référentielle du constituant. Une construction pourvue d'un déterminant ne garantit pas une interprétation en termes de référence actuelle [Milner 1978] ».

[3] Où le GP ([de+*le*] soir) contient un déterminant, sans que ce second N", le GN « le soir », puisse toujours faire l'objet d'une reprise anaphorique et donc constituer un objet de discours.

Rappelons à ce titre que nous avions précédemment relevé quelques remarques allant dans ce sens :

- Le Guern note le caractère particulier du statut « objectal » du second N" lorsqu'il est dans un SP ;
- Berrendonner évoque la possibilité de double interprétation, en termes de « type » et d'« individu », pour certaines unités GN construites avec un SP.

Dans le cadre positionnel, ce qui distingue les compléments aréférentiels de la synapsie et les compléments référentiels du GN, ce n'est pas un critère morpho-lexical (présence/absence de déterminant) mais un critère syntaxique : les deux types de complément, s'ils peuvent avoir la même place en surface, occupent en réalité deux positions distinctes. Sophie David montre, par une série de tests[1], qu'il y a bien deux positions distinctes à droite de la tête nominale. Nous ne rentrerons pas ici dans le détail de sa démonstration.

Cependant, comme nous l'avons déjà noté, si les positions syntaxiques ne sont visibles que lorsqu'elles sont occupées par des « termes », les termes occupants ne montrent pas tout des positions qu'ils occupent. Autrement dit, le caractère aréférentiel des compléments est inégalement et non systématiquement marqué d'un point de vue syntaxique ; c'est là une propriété de la syntaxe même[2] : « La syntaxe ne fournit pas de marque qui puisse fonctionner comme critère diacritique des deux structures dont elle permet pourtant de concevoir les différences.[3] »

On peut tout à fait se doter de critères linguistiques (ici positionnels) dont certains sont propres aux compléments référentiels et d'autres aux compléments aréférentiels, mais encore faut-il que ces critères puissent s'appliquer au matériel linguistique considéré : de ce point de vue, « tous les contextes d'apparition d'une synapsie ne sont pas équivalents[4] ». Certains contextes facilitent plus que d'autres l'identification d'une synapsie.

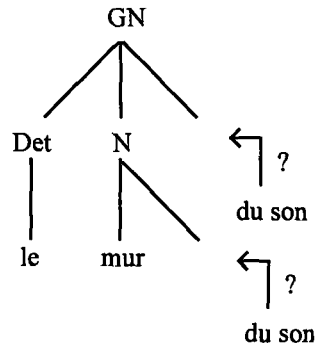
Ainsi, dans « la robe du soir où je l'ai connue », on dispose d'éléments (relatifs au type de la relative) qui permettent de ne pas identifier « robe du soir » comme une synapsie, c'est-à-dire comme une unité susceptible de servir de dénomination de classe (ici la classe des robes du soir) ; le complément « le soir » est ici référentiel. Par opposition, le contexte « la robe du soir de ma mère » présente, lui, une configuration qui permettra avec plus de certitude d'identifier une synapsie, « robe du soir », susceptible de fonctionner, dans ces cas-là, comme dénomination de classe : le complément « le soir » est ici aréférentiel. C'est en cela que la synapsie est une unité construite par la syntaxe ; en aucun cas, elle ne peut être listée dans un lexique : une même forme superficielle sera, selon ce que la syntaxe en fait, une synapsie ou pas, une dénomination de classe ou pas. Les occurrences discursives d'une même forme peuvent avoir des interprétations radicalement différentes.

En disposant d'une théorie de la syntaxe qui en dégage les propriétés, comme celle de « pauvreté », on peut faire le départ entre le rôle discriminant ou non des critères formels et le caractère ambigu ou non du matériel linguistique. C'est pourquoi, dans le cadre positionnel, une structure est qualifiée plus volontiers de « sous-déterminée » que d'ambiguë. Ainsi l'unité « mur du son[5] » peut-elle rester (mais tout

---

[1] David 1993a, p. 152 et suiv.
[2] Propriété dite de « pauvreté » de la syntaxe, Milner 1989, p. 645.
[3] Marandin 1992b, p. 73.
[4] David 1993a, p. 240.
[5] *Ibid.*, p. 230.

dépend de son « entourage » syntaxique) sous-déterminée par rapport à l'interprétation (référentielle ou aréférentielle) :



Si l'identification des unités syntaxiquement construites que sont les synapsies peut être, en raison même de leur nature syntaxique, sous-déterminée, il n'en reste pas moins que l'unité synaptique nous paraît représenter une forme linguistique adéquate à la fonction attendue du descripteur.

B - Synapsie et morphologie du descripteur

Dans le cadre d'une indexation de type explicatif, le descripteur doit être une unité pourvue des propriétés suivantes :
(i) elle doit fonctionner comme une dénomination de classe stable ;
(ii) elle doit permettre de désigner un élément de cette classe.

De ce point de vue, la synapsie apparaît comme une forme possible du descripteur. Elle fait nécessairement partie d'un groupe nominal (c'est une unité lexicale, de type N, située, dans le modèle positionnel, dans la « position noyau » d'un groupe nominal). On sait donc qu'elle peut fonctionner comme un objet de discours (désigner un individu parmi une classe d'individus) : il suffit de la considérer dans le GN où elle apparaît ; en cela, elle répond à la propriété (ii) du descripteur. Par ailleurs, sa construction syntaxique à base de compléments aréférentiels lui permet de fonctionner comme une dénomination de classe : la synapsie est, de ce point de vue, une unité interprétativement saturée. Elle répond en cela à la propriété (i) du descripteur.

Le modèle de la synapsie proposé par Sophie David nous semble introduire un critère pertinent pour notre approche du descripteur : le critère de la position située à droite de la tête nominale permet de distinguer, parmi les différentes unités construites avec un SP, celles qui pourront fonctionner comme désignation de classe stable et celles qui ne le pourront pas.

Si l'on adopte la représentation proposée par Sophie David, la morphologie du descripteur ne se donne plus sous la forme de séquences mais sous forme de configurations de positions syntaxiques, dont certaines pourront rester sous-déterminées. Cependant, on peut montrer que les deux modèles se recouvrent partiellement : la représentation syntaxique de la synapsie couvre à la fois les séquences de type « N + EP » et celles de type « N + SP », mais ne conserve des

séquences « N + SP » que celles où le second groupe nominal ne construit pas d'objets de discours : seules ces séquences constituent des synapsies.

Cependant, l'extraction des synapsies ne saurait se comprendre comme une extraction de descripteurs. La synapsie, étant un individu syntaxique, n'est pas un descripteur : seule une utilisation professionnelle particulière peut la faire fonctionner comme descripteur. De la même façon, une certaine utilisation de la synapsie peut la faire fonctionner, par exemple, comme une « unité terminologique[1] ».

Ainsi voit-on qu'un même type d'unité, doté de la propriété de désigner « rigidement » dans un cercle donné de locuteurs, peut être utilisé dans des contextes de communication opposés : à la « communication spécialisée » propre à la terminologie s'oppose la « communication vulgarisante » propre à la documentation. Utilisées comme des termes, les synapsies permettent de circonscrire des domaines spécialisés ; utilisées comme des descripteurs, elles permettent d'établir des ponts entre des domaines de spécialité et d'évoluer au sein d'une formation discursive.

Si la nature linguistique du descripteur joue un rôle prépondérant en indexation, seule une stratégie d'exposition proprement documentaire, fondée sur une stratégie spécifique d'exploration des sources, peut garantir l'efficacité du traitement réalisé. Le descripteur n'est donc pas une unité linguistique d'un type particulier : il emprunte sa forme à la langue même. Reste, pour l'indexation, à l'exploiter dans un cadre discursif adéquat.

Sur ce point, en présentant les systèmes automatisés dont ont fait l'objet les deux représentations linguistiques proposées, on pourra indiquer brièvement, selon le point de vue que l'on privilégie (indexation/recherche documentaire), comment il est possible de tirer partie des descripteurs quand ils sont unités de discours : groupes nominaux complexes ou synapsies.

III.3 - Extraction automatique d'unités de discours

Pour une description technique des systèmes conçus à partir des représentations linguistiques présentées ci-dessus, nous renvoyons aux textes suivants[2] :

- pour le modèle de Michel Le Guern, on peut consulter sur le système Sydo[3] : Berrendonner, Bouché, Le Guern, et *al.* [1980], Metzger [1988], Le Guern [1989] ;
- pour le modèle proposé par Sophie David, on peut consulter sur le système Termino[4] : David et Plante 1990a et 1990b, David 1993a.

Dans ce paragraphe, nous indiquerons quelques particularités des systèmes automatiques proposés en insistant sur le type de traitement linguistique qu'ils

---

[1] Perron 1988. Le Guern (1989 par exemple) a également rapproché indexation et terminologie sur la base de leurs types d'unité.

[2] Bibliographie partielle, qui peut être complétée par les références citées dans ces textes.

[3] Nous appelons « système Sydo » ou « Sydo » tout court l'analyseur morpho-syntaxique développé par l'équipe SYDO.

[4] Termino a été conçu par S. David, L. Dumas, J.-M. Marandin, A. Plante et P. Plante, avec la collaboration de D. Perras et I. Winter.

réalisent et sur l'usage professionnel qui peut être fait de ces traitements de nature linguistique. On présentera brièvement :

- les caractéristiques du traitement linguistique effectué (III.3.1) ;
- les usages possibles selon le point de vue privilégié (indexation/recherche documentaire) § III.3.2.

III.3.1 - SYSTÈMES LINGUISTIQUES D'EXTRACTION D'UNITÉS DE DISCOURS

Ce qui permet de pouvoir utiliser Sydo ou Termino dans un cadre professionnel, c'est, d'une certaine façon, la certitude que les deux systèmes automatisés extraient bel et bien des unités de discours et non des chaînes de caractères. Ce qui paradoxalement garantit la fiabilité des unités repérées, c'est le fait que les deux systèmes ne cherchent pas à repérer l'unité conceptuellement hétérogène qu'est le descripteur mais des unités linguistiques conceptuellement « homogènes » : la distinction entre propriétés de langue et usage documentaire des propriétés de langue est, en matière de traitement automatique, tout particulièrement cruciale.

De ce point de vue, les deux systèmes ont en commun de ne disposer que d'un savoir morpho-syntaxique : aucune connaissance de type sémantique et encore moins pragmatique (relevant de la pratique documentaire) n'est nécessaire à l'extraction d'unités de discours utilisables en indexation. Si ce savoir morpho-syntaxique est différent dans les deux systèmes, un même type d'approche linguistique des textes est à l'œuvre. En effet, on ne peut envisager une extraction d'unités de discours sans modules préalables de traitement capables de :

- segmenter un texte en « phrases » et en « mots » ;
- lemmatiser[1] les « mots » extraits ;
- effectuer une représentation syntaxique des énoncés[2].

Aucune de ces trois opérations n'est triviale. Par exemple, le seul examen des signes de ponctuation ne garantit pas un découpage en phrases (si le point (.) semble fonctionner comme le marqueur par excellence des fins de phrases, comment le traiter quand il apparaît dans un sigle (S.N.C.F.) ?). De même, la segmentation en « mots » ne peut se fonder sur le seul espace entre chaînes de caractères (les suites avec apostrophe risquent d'être toujours comptées pour une unité alors qu'il peut y en avoir plusieurs : « l' » et « enfant » dans la suite « l'enfant » *versus* « aujourd'hui ») ; etc.[3]

Les deux systèmes mettent en œuvre des procédures différentes pour traiter ces trois aspects préalablement nécessaires à l'extraction proprement dite des unités de discours ; nous renvoyons, pour le détail des choix, aux références bibliographiques précédemment données.

L'extraction automatique d'unités de discours réalisée avec les systèmes Sydo et Termino ne présentent aucun des désavantages habituellement avancés par les

---

[1] La lemmatisation consiste à rapporter la diversité flexionnelle à une forme canonique. La forme canonique pour un adjectif est la forme au masculin singulier (belles -> beau) ; celle du nom est le singulier (chevaux -> cheval) ; celle du verbe, l'infinitif (mangeait -> manger).

[2] Les textes sont, dans leur intégralité, soumis à une analyse morpho-syntaxique. Pour une critique de la notion d'analyse syntaxique « superficielle » (ne visant que le seul repérage des formes « syntagme nominal »), voir David 1993a, p. 192-218.

[3] Pour une revue de ces problèmes, propres au domaine du traitement automatique des langues, voir, par exemple, Fuchs 1993.

professionnels comme arguments contre l'indexation-extraction[1]. *Les unités dépistées sont toutes des unités bien formées du point de vue d'un modèle explicite de la langue. Comment les utiliser d'un point de vue documentaire ? Nous présentons ci-après quelques pistes d'exploitation de ces procédures d'extraction automatique, selon que l'on privilégie le point de vue de l'utilisateur ou celui de l'indexeur.*

III.3.2 - UTILISATION DOCUMENTAIRE DES UNITÉS DE DISCOURS

Le modèle par lequel Michel Le Guern propose de définir le descripteur nous paraît de nature à faire apparaître une différence entre le descripteur des indexeurs et le descripteur des utilisateurs. Selon que le produit de l'indexation-extraction est directement utilisé dans un système d'information ou pas, la morphologie du descripteur pourra être différente : le syntagme nominal constitue par excellence la forme pertinente du descripteur dans un système d'information ; la synapsie constitue, selon nous, la forme adéquate du descripteur pour les indexeurs. Cependant, dans les deux cas, c'est sur la base d'un même type de propriété (propriété d'emboîtement) que les unités linguistiques pourront être « documentairement » utilisées. C'est en outre cette même propriété qui peut permettre de spécifier l'usage documentaire qu'il est possible de faire des descripteurs de forme simple.

A - Principe d'emboîtement

On entend par « principe d'emboîtement » l'inclusion d'unités dans d'autres unités plus larges, que celles-ci soient des groupes nominaux ou des synapsies :

- exemple de syntagmes emboîtés[2] : « la représentation des salariés » est un syntagme que l'on dira englobant dans la mesure où il comprend le syntagme « les salariés », que l'on dira englobé ;
- exemple de synapsies emboîtées[3] : « traitement sans procédure de chargement initial » sera qualifié de synapsie englobante puisqu'elle contient la synapsie, dite alors englobée, « procédure de chargement initial ».

Là encore, on remarquera que le découpage des unités englobantes s'effectue sur une base linguistique permettant de garantir la constitution d'objets de discours. Pour montrer la différence entre un découpage linguistique et un découpage aléatoire de chaînes de caractères, considérons les deux exemples suivants :

- Dans la séquence (i) « les conditions de l'amélioration de la représentation des salariés dans les PME », l'analyse du groupe nominal effectué par Sydo ne produira pas des séquences de type « les conditions de l'amélioration » par exemple, ou encore une séquence telle que « l'amélioration de la représentation ». Non que ces segments ne puissent, dans d'autres occurrences, constituer des syntagmes nominaux, et peut-être des descripteurs, mais dans le cadre particulier de la séquence (i), ces segments ne constitueront jamais des objets de discours et ne seront donc jamais proposés comme tels ;

- De même, Termino traitant l'énoncé (ii) suivant « J'étudie un système de dépistage d'unités terminologiques », ne produira jamais, parce que ce sont

---

[1] Voir précédemment § I.2.1.
[2] Repris de Le Guern 1991a.
[3] Repris de David 1993a, p. 242.

pas dans ce contexte des synapsies, des séquences de type « système de dépistage » ou encore « dépistage d'unité ». Ces séquences ne répondent à aucun critère de découpage linguistique, à aucune des propriétés syntaxiques des synapsies ; à ce titre, elles ne pourront jamais constituer des dénominations de classe.

Sans pouvoir rentrer ici dans le détail des règles spécifiquement établies dans les systèmes Sydo et Termino, nous voudrions simplement insister sur le fait que le repérage et le découpage linguistiques effectués par les deux systèmes, s'ils peuvent être sous-déterminés dans le cas de Termino ou jugés trop « bruyants » dans le cas de Sydo, garantissent que les unités de discours extraites peuvent réellement fonctionner comme dénominations de classe. L'usage viendra confirmer ou infirmer une telle potentialité, mais ce n'est pas à partir de l'usage que l'on peut identifier les unités utilisées comme dénominations de classe.

Là encore apparaît cette nécessité, paradoxale, de préférer une indexation-extraction automatique des unités de discours à une indexation manuelle : la seconde, qui mêle de façon intrinsèque les deux caractéristiques (langue et usage) des objets de discours, n'est pas toujours à même de repérer de « nouvelles » dénominations[1]. La tentation de reconnaître dans de nouvelles unités d'« anciennes » (celles que l'on connaît déjà) est ici fatale. C'est, selon nous, de cet ordre que relève le glissement interprétatif à l'œuvre dans les indexations par « Bernard Tapie » de l'article consacré au redressement judiciaire de ses sociétés[2]. Sur ce point, l'article, soumis à l'analyse linguistique de Termino[3], révèle l'existence de deux unités distinctes : « redressement judiciaire », qui constitue une dénomination de classe autonome, et « redressement judiciaire à titre personnel », qui constitue une autre dénomination de classe. Il s'agit là de deux synapsies distinctes, non emboîtées l'une dans l'autre, susceptibles de construire deux objets de discours différents. À partir de ces deux expressions, comprises comme différentes, il est possible de voir que, dans l'article en question, il est fait mention de deux types de jugement distincts : l'un concerne le patrimoine personnel de Bernard Tapie (cet aspect n'est que peu développé dans l'article), l'autre concerne la gestion de ses sociétés (et constitue le cœur de l'article). Sur ce point, il nous semble qu'une extraction automatique se révèle nécessairement plus fiable qu'une indexation manuelle, pour peu qu'on prenne la mesure de la spécificité des objets textuels et de là, que l'on considère comme première la nécessité de considérer les descripteurs comme unités extraites de ces textes-là.

Le type d'unités extrait par Sydo ou Termino, englobé et/ou englobant, porte donc une structuration interne qui permet de les utiliser d'au moins deux façons :

- *en recherche d'information, le parcours dans un texte est facilité par l'exploitation de ce principe d'emboîtement (B) ;*
- *en indexation, la structuration des synapsies peut permettre aux indexeurs d'organiser les accès qu'ils proposent aux textes (C).*

---

[1] Le caractère « nouveau » est relatif, en indexation, à la connaissance spécifique à chaque indexeur. Chaque désignation relève, comme nous l'avons vu, d'un cercle de locuteurs particulier, le plus souvent restreint. Un indexeur ne peut connaître toutes les dénominations mises en œuvre dans les textes qu'il indexe et, comme un texte ne met que rarement en jeu qu'un seul domaine de spécialité, confier l'indexation à des « spécialistes » ne constitue pas forcément la meilleure des solutions.

[2] *Supra.*

[3] L'analyseur morpho-syntaxique Sydo aurait pu, tout autant, révéler cette différence, mais nous ne l'avions pas à disposition pour effectuer des tests.

Dans les deux cas, c'est sur la base du principe d'emboîtement que les descripteurs de forme « simple » pourront trouver une pertinence.

B - Principe d'emboîtement en recherche documentaire

Les systèmes d'information conçus par l'équipe Sydo reposent sur la structuration des groupes nominaux[1], cette structuration garantissant que l'accès à un texte par une forme superficiellement simple conduit bien à un objet de discours. Reprenons sur ce point l'exemple présenté dans Le Guern [1991a].

À partir du syntagme « les conditions de l'amélioration de la représentation des salariés dans les PME », on peut dégager cinq syntagmes nominaux répartis sur quatre niveaux :
(4) les conditions de l'amélioration de la représentation des salariés dans les PME
(3) l'amélioration de la représentation des salariés dans les PME
(2) la représentation des salariés dans les PME
(1) les salariés
(1) les PME

En exploitant le principe d'emboîtement des syntagmes nominaux, un utilisateur peut tout à fait choisir d'« entrer » dans une collection documentaire par le descripteur de forme simple « les salariés » mais ce choix le mènera à des unités de discours de type (2), (3) ou (4) qui, d'emblée, en précisant le domaine d'interprétation de l'unité « les salariés », pourront être de nature à le renseigner : l'objet « les salariés » ainsi traité dans ce texte m'intéresse-t-il ou non ?

Par ailleurs, une même entrée de type « descripteur simple » comme « les PME » peut mener à un autre descripteur complexe, référentiellement synonymique du premier, par exemple « la représentation des salariés *au sein des* PME » : les relations de coréférence entre syntagmes nominaux peuvent ainsi être captées par le biais du principe de l'emboîtement.

Il est donc possible, dans le cadre de la recherche d'information, de concevoir le descripteur sous une forme simple. Reste que, compte tenu du statut objectal particulier des groupes nominaux englobés (voir précédemment), on devra privilégier l'accès par des formes « complexes » interprétativement autonomes et sélectionner, selon des critères qu'il reste à définir précisément, les descripteurs simples susceptibles de constituer des accès pertinents. L'exploitation documentaire des indices linguistiques est en effet loin d'être triviale. Sur ce point, si l'exploitation des structures attributives dans un texte présente une possibilité d'étendre la structuration des groupes nominaux, Michel Le Guern[2] montre bien, sur ce plan encore, la nécessité d'un examen attentif des individus linguistiques en jeu.

---

[1] Le Guern 1991a, p. 34 : « En soi, la liste de tous les syntagmes nominaux du corpus, accompagnés pour chacun de la liste des références de ses occurrences, est déjà utile. Mais, *pour une plus grande efficacité de l'outil d'interrogation, il convient de structurer cet ensemble de syntagmes nominaux, en se servant d'abord des relations d'emboîtement* ». C'est nous qui soulignons.
[2] Sur ce point, voir Le Guern 1991b.

C - Principe d'emboîtement en indexation

Rappelons que l'emboîtement des synapsies entre elles, qui repose sur un principe de construction linguistique (C2), se distingue de l'inclusion de la synapsie dans un groupe nominal, inclusion qui met en jeu deux « niveaux » distincts (C1). Nous précisons ci-après la différence par quelques exemples extraits du traitement linguistique que Termino a réalisé à partir du numéro du *Monde* [1] qui nous sert de référentiel expérimental dans cette recherche.

C1 - Inclusion de la synapsie dans un groupe nominal

Comme nous l'avons précédemment noté, l'un des traits définitoires de la synapsie est d'être incluse dans la position noyau du groupe nominal, lui permettant ainsi d'apparaître au sein de référents discursifs différents. Ainsi la synapsie « usage de stupéfiants » apparaît-elle sous les « angles » discursifs suivants dans l'article[2] du *Monde* que nous avons soumis à Termino :
(i) la dépénalisation de l'*usage de stupéfiants*
(ii) la condamnation pour violence avec armes et *usages de stupéfiants*
(iii) le simple *usage de stupéfiants*

De ce point de vue, la synapsie fonctionne comme le terme simple « les salariés » de l'exemple précédemment présenté. C'est de ce point de vue que, nous semble-t-il, Michel le Guern considère que la synapsie correspond, dans son modèle, au niveau N, qu'il conviendrait d'appréhender, pour l'indexation, au niveau N'[3].

Comme dans le paragraphe précédent, la découverte d'unités comme (i) à partir de la synapsie « usage de stupéfiants » peut permettre à un utilisateur de poursuivre son exploration, en sélectionnant par exemple les synapsies qui ont pour tête « dépénalisation » : il sera alors conduit vers des formes de types « dépénalisation à demi-mot », « dépénalisation de l'usage de drogue », etc.

De ce point de vue, l'exploitation documentaire des synapsies rejoint, dans le principe, celle qu'un utilisateur peut mener à partir des syntagmes nominaux. L'emboîtement des synapsies entre elles présente, lui, la possibilité pour les indexeurs d'effectuer un classement préalable des unités, de nature à aider la recherche documentaire. Le fait que les synapsies soient des unités interprétativement saturées permet de pouvoir les manipuler sans qu'une interprétation discursive ne soit nécessairement en jeu. Par opposition, la manipulation des groupes nominaux semble plus naturellement revenir aux utilisateurs qui organisent alors leur parcours discursif du point de vue de leur problématique personnelle.

C2 - Emboîtement des synapsies

L'emboîtement des synapsies relève du mode de construction qui les caractérise en propre.

---

[1] *Le Monde* du 1er décembre 1994.
[2] *Ibid.*, p. 11.
[3] Communication personnelle : « J'aurais tendance à penser que ce n'est pas la synapsie pure qui fournirait un descripteur, mais sa projection en logique extensionnelle ».

Comme le fait remarquer Benveniste, la « nouvelle forme de composition nominale » que sont les synapsies présente cette particularité d'être très productive. Cette spécificité, qui la distingue des autres modes de composition, tient à sa structuration syntaxique : « C'est là un procédé qui contraste avec la composition traditionnelle par la facilité et l'ampleur de ces réalisations. Alors que la composition, en français, atteint très vite ses limites, et que les composés se forment à un rythme lent et pour ainsi dire par cooptation individuelle (on voit apparaître les premiers spécimens d'une série nouvelle en cosmo- avec *cosmonaute*, *cosmodrome*), la synapsie prodigue sans trêve ses créations. Tous les vocabulaires techniques y font appel, et d'autant plus aisément qu'elle seule permet la spécification détaillée du désigné, et la classification des séries par leur trait distinctif. Son extrême flexibilité paradigmatique fait de la synapsie l'instrument par excellence des nomenclatures.[1] »

En effet, on peut relever, dans les domaines dits techniques (par exemple l'informatique, pour les exemples ci-dessous), la constitution de paradigmes qui repose sur l'emboîtement de synapsies, ainsi de la série :
(i) base de données
(ii) gestion de *bases de données*
(iii) système de *gestion de bases de données*

On peut généraliser ce principe de constitution de paradigmes en regroupant les synapsies sur la base de la synapsie qui y est englobée. Ce type d'utilisation du principe d'emboîtement spécifique à la « nouvelle forme de composition nominale » qu'est la synapsie est régulièrement sollicitée en terminologie[2]. On dispose ainsi de structuration du type suivant, réalisé à partir de la synapsie englobée « carte à mémoire[3] » :

application de la carte à mémoire
coût de la carte à mémoire
départ de carte à mémoire
développement de la carte à mémoire
domaine de la carte à mémoire
émergence de la carte à mémoire
essor de la carte à mémoire
fabricant de carte à mémoire
industrie de la carte à mémoire
marché de la carte à mémoire
pénétration de la carte à mémoire
percée de la carte à mémoire
possibilité de la carte à mémoire
projet de la carte à mémoire
projet de carte à mémoire multi-services
technologie de la carte à mémoire
terminal de la carte à mémoire
vidéotex par carte à mémoire[4]

À partir de cette liste, on peut structurer les synapsies, par exemple, par « facettes » en créant des sous-ensembles de synapsies dont la tête est de type : « application », « coût », « industrie », « technologie », par exemple.

Il nous semble que l'on dispose là de quoi pouvoir mener une véritable réflexion sur l'organisation des accès documentaires à proposer à des utilisateurs, sans que l'on ait à connaître leurs « intérêts » ou encore leur « besoin » d'information : on peut, à partir des textes eux-mêmes et de leurs unités de discours, organiser un espace de recherche documentaire cohérent.

---

[1] Benveniste 1974 [1966], p. 174.
[2] Sur ce point, voir Perron 1988 et 1991.
[3] Les synapsies présentées ici sont issues de l'analyse linguistique que Termino a effectuée sur l'article *La carte à mémoire s'utilise partout, par tous* paru dans *Informatique & Bureautique*, mars 1990, p. 16, 18-23, repris de Perron 1991, p. 749
[4] Repris de Perron 1991, p. 749.

Ce qui est ici valable pour des domaines dits techniques nous semble tout à fait réalisable sur des textes que l'on qualifie volontiers de « généraux » ou d'« encyclopédiques ». Dans cette perspective, nous avons soumis le numéro du *Monde* qui nous sert d'exemplaire de travail au logiciel Termino : notre but était simplement de voir s'il était possible d'organiser les synapsies en « paradigmes ». Notre expérimentation ne saurait valoir d'exemple d'indexation automatique : elle ne fait appel à aucune autre source. Or l'indexation ne vaut que par le choix et l'organisation des sources qu'elle effectue, que par les liens entre documents qu'elle rend possible.

Du point de vue de la seule possibilité d'une mise en paradigmes, l'expérience menée montre qu'une structuration des données sur la base d'un emboîtement de synapsie est possible ; on peut ainsi constituer :

- à partir de la synapsie « redressement judiciaire », le regroupement « procédure de *redressement judiciaire* » et « mise en *redressement judiciaire* » ;
- à partir de la synapsie « cumul des mandats », le regroupement « limitation du *cumul des mandats* » et « restriction du *cumul des mandats* » ;
- à partir de la synapsie « construction européenne », le regroupement « avenir de la *construction européenne* » et « poursuite de la *construction européenne* ».

L'examen des synapsies emboîtées nous semble tout à fait indispensable. Cet examen peut permettre aux indexeurs de décider ce qu'ils choisissent de présenter aux utilisateurs : toutes les synapsies, englobées et englobantes, uniquement les synapsies englobées, accompagnées de leurs regroupements, uniquement les synapsies les plus englobantes, etc., avec toujours la certitude qu'il s'agit bien là d'unités de discours.

Le principe d'emboîtement de la synapsie peut être encore utilisé d'une autre façon, susceptible elle aussi d'aider le travail de structuration des descripteurs réalisé par les indexeurs. Comme précédemment, on peut utiliser les descripteurs de forme complexe pour déterminer des descripteurs de forme simple susceptibles d'être des accès documentaires pertinents.

Sur ce point, le logiciel Termino propose de distinguer, parmi l'ensemble des noms « simples » d'un texte, ceux qui présentent la conjonction des deux propriétés suivantes :

- les unités nominales doivent apparaître, dans le corpus soumis à l'analyse, au moins une fois dans une synapsie, soit en position de « tête » soit en position de « complément » ;
- elles doivent en outre apparaître, au moins une fois, en position de tête dans un groupe nominal dépourvu de complément.

L'heuristique ici mise en place a pour but de chercher à capter les formes nominales susceptibles de fonctionner, d'un point de vue interprétatif, comme termes « génériques ». Sous cet angle, l'unité nominale « redressement » apparaît, dans notre corpus, comme élément de cette sous-classe de nom simple : l'unité nominale apparaît en position de tête dans les synapsies « redressement fiscal », « redressement judiciaire », « redressement judiciaire à titre personnel » et « redressement judiciaire de la quasi-totalité des sociétés ». L'unité nominale apparaît par ailleurs dans un groupe nominal comme unité de discours

« autonome[1] ». À ce titre, elle remplit les deux conditions requises. Sur un plan interprétatif, on peut en effet considérer que les quatre synapsies constituées avec la tête nominale « redressement » constituent des « types » de redressement : à ce titre, on peut considérer que le terme « redressement » est un terme générique.

Il importe de relever sur ce point que toutes les têtes nominales ne sont pas susceptibles de tels emplois génériques. Comme le montre Michel Le Guern[2], on ne saurait reconnaître la même unité « trompe » dans les deux expressions suivantes : « trompe d'Eustache » et « trompe de chasse », et, précise-t-il, « trompe » ne peut constituer un terme, d'un point de vue terminologique ; il ne peut pas non plus, d'un point de vue documentaire, constituer un descripteur. Là encore, l'exploitation documentaire d'indices linguistiques exige un examen attentif des individus linguistiques en cause.

Autrement dit, si les descripteurs peuvent être de forme simple, comme nous avons essayé de le montrer, ce ne peut être que dans le cadre de configurations très particulières, sans doute très délicates à manipuler dès lors que le volume d'une collection documentaire atteint un volume important : il s'agit bien là d'exceptions, comme l'a montré Le Guern[3], et la forme normale du descripteur reste celle d'une forme « complexe ».

Nous voulions dans le dernier paragraphe de ce chapitre présenter des systèmes d'analyse linguistique permettant de réaliser une extraction automatique d'unités de discours : groupes nominaux ou synapsies. Nous voulions également montrer que la seule extraction automatique d'unités de discours, si elle nous paraît indispensable, ne saurait être assimilée à l'indexation documentaire :

- *le traitement des sources en amont constitue une étape indispensable à la viabilité de l'opération d'extraction proprement dite ;*
- *une réflexion sur l'utilisation documentaire des mots extraits du discours est tout autant indispensable. On peut choisir de se placer dans le cadre d'un système d'information (cas du système Sydo) ; on peut choisir de rester dans le domaine de l'indexation et, à ce niveau-là, réfléchir à la structuration des accès qu'il convient de donner (cas de l'utilisation du logiciel Termino pour l'indexation). Dans les deux cas, il s'agit d'exploiter des propriétés de langue dans le cadre d'une pratique professionnelle particulière.*

Tout comme dans le chapitre IV, où nous avons essayé de circonscrire la zone d'intervention propre à la pratique dans la constitution des documents, nous avons cherché ici à montrer la nécessité qu'il y a de distinguer les niveaux dans toute approche du descripteur : parce que le descripteur est un individu linguistique en même temps qu'un des « outils » de l'indexeur, les deux dimensions doivent être maintenues séparées. On a également tâché de montrer que ces deux « dimensions » étaient hiérarchisées : la dimension linguistique contraint la dimension documentaire. En cela, la pratique professionnelle de l'indexation doit s'établir sur la base des propriétés linguistiques des textes eux-mêmes et des unités linguistiques elles-mêmes : les conditions d'existence des textes (chapitre IV), ainsi que le mode de fonctionnement des unités en discours, doivent être considérés préalablement à toute utilisation documentaire.

---

[1] Dans le groupe nominal « la perspective d'un quelconque *redressement* », *Le Monde*, 1/12/1994, p. 24.
[2] Le Guern 1989, p. 343.
[3] Le Guern 1984, p. 167.

IV - Conclusion du chapitre

Nous avons, dans ce chapitre, abordé l'indexation sous l'angle du descripteur, c'est-à-dire dans sa phase de détermination de descripteurs, phase à laquelle elle est le plus souvent réduite, du moins dans les approches classiques. Rappelons que, dans notre approche, la détermination des descripteurs ne constitue qu'un aspect, certes le plus visible, de l'indexation, dont nous pensons cependant que l'essentiel concerne l'organisation (et le choix) des documents qu'elle met à disposition.

Dans le premier paragraphe de ce chapitre est apparue la nécessité de disposer d'un cadre théorique qui nous permette de poser la problématique du descripteur en indexation. En effet, nous avons pu constater que :

- les approches classiques ne permettent pas de définir la morphologie du descripteur. La forme linguistique du descripteur y est jugée secondaire (c'est par commodité qu'est retenue la forme substantive) dans la mesure où la problématique privilégiée est celle du langage documentaire et non celle du descripteur à proprement parler : le descripteur n'est approché que par le rôle qu'il tient dans un langage documentaire. Or, nous avons précédemment pu montrer que la problématique des langages documentaires, qui considère l'indexation dans sa seule dimension lexicale, n'était pas de nature à rendre compte de la fonction des mots en indexation (leur rôle référentiel et par suite leur rôle en discours). La prédominance de la problématique des langages documentaires au détriment d'une problématique spécifique au descripteur nous a fait définitivement renoncer à considérer le modèle normatif comme un modèle de référence : nous avons à ce titre proposé d'autres modes d'approche de l'indexation et de la recherche documentaires ;

- l'absence de problématique spécifique au descripteur conduit en outre à des évaluations erronées de l'indexation-extraction. On a ainsi pu montrer que les professionnels sont souvent conduits à confondre la procédure d'extraction elle-même avec le type d'unité extrait des textes et proposé comme descripteur : là encore, l'absence de référentiel spécifique ne permet pas toujours de distinguer les éléments en jeu.

Le cadre théorique retenu pour étudier la problématique du descripteur repose sur des propriétés linguistiques, qui constituent une partie des fondements théoriques de l'indexation : du point de vue de la théorie linguistique, l'information peut se définir comme la construction d'un thème de discours.

Sur cette base se dégagent deux types d'indexation possibles :

- un type d'indexation que l'on a appelé « interprétatif » et qui consiste à disposer l'espace de thématisation du côté des indexeurs ; ce type d'indexation qui contredit les propriétés que l'on peut attribuer au document (usages multiples, détournements, etc.), ne nous semble pas adéquat ;

- un type d'indexation que l'on a qualifié d'« explicatif », dans lequel l'espace de thématisation est situé du côté des utilisateurs : ce type d'indexation semble plus conforme aux fonctions attendues du document. Ce type d'indexation présente la particularité de se situer en deçà de l'interprétation et, en cela, il se révèle d'autant plus fiable qu'il est effectué de façon automatique : nous arrivons là à une conclusion qui pourra paraître bien étrange.

C'est en privilégiant le point de vue de l'indexation explicative qu'apparaît la nécessité d'approfondir la manière dont se construit le thème de discours : la construction d'un thème de discours s'effectue par le biais d'une série de saisies de différents référents discursifs. Une façon de modéliser ce processus est de recourir au modèle des chaînes de référence proposé par Chastain. Dans ce cadre, on peut étudier la morphologie du descripteur, en examinant d'abord le fonctionnement logique des unités et ensuite leurs propriétés linguistiques.

L'examen mené d'un point de vue logique, dans le second paragraphe de ce chapitre, a montré que le candidat « nom propre » ne constituait pas, contre toute attente, un bon candidat-descripteur, du moins dans le cadre d'une indexation explicative : les atouts de sa rigidité (identité formelle = identité référentielle) se muent en obstacles dans le cadre de la construction des chaînes de référence en indexation. En revanche, les emplois rigides des descriptions définies représentent, eux, les emplois typiques des descripteurs, leur permettant d'être à la fois des dénominations stables de classe et des éléments d'identification singulière de référents discursifs, et à terme, de référents mondains.

L'étude linguistique du descripteur, proposée dans le dernier paragraphe de ce chapitre, a permis de distinguer, dans le cadre d'un modèle général qui spécifie le descripteur comme unité de discours, deux « états » du descripteur selon le point de vue privilégié. Le descripteur de la recherche documentaire n'est pas tout à fait le même que le descripteur de l'indexation documentaire, ce qui souligne, une fois encore, qu'indexation et recherche ne constituent pas des opérations similaires, symétriques, dans lesquelles les mêmes formes pourraient être indistinctement employées. Le descripteur utilisé dans une recherche documentaire correspond au syntagme nominal de forme « complexe » ; le descripteur capté au niveau de l'indexation elle-même se présente, de façon privilégiée, sous la forme d'une synapsie : la relation aux discours n'est pas tout à fait la même. Ces deux « états » du descripteur sont intimement liés, la synapsie étant, par sa construction syntaxique, insérée dans un groupe nominal ; mais il nous semble important de suggérer ici une différence, de façon à permettre aux indexeurs d'une part et aux concepteurs de systèmes d'information d'autre part de disposer d'un type d'unité réellement conforme au niveau d'intervention, de travail, dans lequel ils se situent.

CONCLUSION DE LA DEUXIÈME PARTIE

Notre contribution aux fondements théoriques de l'indexation se signale par la mise en perspective d'un modèle d'utilisation de la langue qui repose sur un modèle de fonctionnement linguistique.

Rappelons que nous avons proposé de faire reposer le mécanisme de l'indexation sur le fonctionnement linguistique du thème discursif. Cet objet linguistique permet d'unifier les différentes propriétés de langue qui nous semblent être à l'œuvre en indexation : la signification lexicale, la synonymie référentielle, la construction de la référence, la « rigidité » désignative. À partir de cet objet et des propriétés qui le sous-tendent, nous avons cherché à disposer d'approches formelles permettant de pouvoir les rendre opératoires dans le cadre des problématiques documentaires.

Il nous paraît en effet essentiel de fonder les pratiques d'indexation sur les spécificités, les propriétés mêmes, des objets qu'elle manipule. Sur ce point, il nous semble indispensable de considérer en indexation :

(i) les conditions d'existence des textes ; l'approche proposée par Foucault constitue une piste pour repenser la notion de politique d'acquisition et les règles de sélection des sources ;

(ii) les conditions d'existence des objets de discours ; sur ce point, le modèle établi par Kripke fournit un éclairage précieux pour spécifier la problématique de la stabilité référentielle en indexation ;

(iii) les conditions d'existence de la référence discursive ; c'est ici le modèle de Chastain qui fournit le cadre général d'une approche explicite et autonome de la morphologie du descripteur.

Dès lors que l'on prend en compte la langue telle qu'elle se donne (le fonctionnement interprétatif des textes comme celui des unités qui les constituent), l'indexation ne se laisse plus définir sous l'angle de la seule détermination de descripteurs. À cette approche strictement lexicale, on est amené à substituer une approche essentiellement discursive : l'indexation se conçoit principalement sous l'angle de l'espace documentaire qu'elle constitue. En effet, il apparaît que le repérage proprement dit des accès documentaires devrait être idéalement effectué

de façon automatique, ce type de repérage présentant la particularité de se situer en deçà de l'interprétation.

Dès lors, le travail des indexeurs se trouve redéfini ; on peut distinguer :

- un ensemble de tâches concernant la constitution des collections documentaires. Nous n'avons pas spécifié cet aspect du travail des indexeurs autrement que par la proposition d'un ancrage théorique. Toute une étude reste à faire pour spécifier une méthode de constitution de l'interdiscours en indexation ; sans doute l'apport des domaines de recherche historique et sociologique est-il ici souhaitable ;

- un ensemble de tâches concernant l'exposition des collections documentaires et, dans les contextes informatisés, l'ergonomie des systèmes d'information. Là encore, nos propos restent suggestifs. Nous n'avons que suggéré la constitution de regroupements textuels sur la base de la terminologie des textes ; sur ce point, les problématiques de la terminologie et de la vulgarisation scientifique pourraient être profitablement explorées. Le travail sur les accès documentaires à mettre à disposition des usagers, au double niveau de l'accès à un domaine et de l'accès à une formation discursive, n'a été, lui aussi, qu'esquissé : les procédures adéquates d'articulation des deux niveaux nous semblent pouvoir être pensées dans le cadre des problématiques de l'ergonomie cognitive.

On le voit : si cette recherche permet de dégager des fondements de l'indexation du point de vue de la théorie linguistique, elle ne permet pas pour autant de constituer, à elle seule, une formalisation complète du processus de l'indexation. Reste que notre approche linguistique des faits d'indexation nous a permis d'opérer, nous semble-t-il, un déplacement majeur dans la compréhension du mécanisme documentaire.

Si la notion de thématisation est commune aux approches classiques et à l'approche linguistique que nous proposons, elle se laisse différemment décrire :

- dans l'approche classique, le thème est un fait de production et met en avant une approche intradiscursive : c'est le thème tel qu'un indexeur le dégage à la lecture d'un document qui constitue un descripteur ;

- dans l'approche que nous proposons, le thème est un fait d'interprétation et met en avant une approche interdiscursive. La thématisation se situe du côté des utilisateurs. Ce choix est possible car la connaissance des mécanismes langagiers permet de pouvoir contraindre les parcours interprétatifs : l'« ambiguïté » de la langue ou encore l'hétérogénéité des textes ne sont plus perçues comme des obstacles que doit détourner l'indexation. Au contraire, la langue apparaît comme un allié précieux, avec lequel les indexeurs doivent pouvoir coopérer.

CONCLUSION GÉNÉRALE

L'indexation documentaire est le plus souvent appréhendée par les professionnels de l'information et de la communication sous la seule dimension instrumentale. C'est en se focalisant sur la question « à quoi sert-elle ? » (en l'occurrence à la recherche d'information) que l'indexation se définit, se laisse décrire et évaluer.

Le choix d'un tel angle d'approche se comprend aisément dans le cadre des pratiques professionnelles où dominent les exigences de performance. Toutefois, il se révèle réducteur dès lors que l'indexation devient l'objet d'autres problématiques, notamment celles de l'évolution des technologies de l'information. Le point de vue « instrumental » ne permet pas alors de discuter l'indexation dans ses spécificités. Il conduit de proche en proche à la faire apparaître comme une simple « technique », parmi d'autres techniques possibles : l'objet « indexation » proprement dit tend à disparaître des problématiques documentaires ; du moins, il n'est plus considéré comme central.

Mettant à distance la finalité ainsi entendue de l'indexation, cette recherche s'est intéressée à dégager les spécificités, les caractéristiques, les propriétés de l'indexation. Pour ce faire, elle a cherché à établir les fondements théoriques de l'indexation. Ceux-ci ont été appréhendés partiellement, du point de vue d'une théorie linguistique.

I - Le point de vue linguistique sur l'indexation : une tentative de distinction

Le point de vue linguistique sur l'indexation permet de dégager les représentations sous-jacentes de la langue et du langage à l'œuvre dans les pratiques professionnelles. Apparaissent alors des confusions entre faits de langue et faits d'indexation. Des principes propres à l'interprétation des discours comme à celle des unités lexicales se donnent comme des objectifs singuliers de l'indexation.

Ainsi la notion de « représentation du contenu d'un texte par des mots » devient-elle définitoire du processus de l'indexation, engageant la construction

d'instruments spécifiques comme les langages documentaires, alors que le point de vue linguistique montre qu'il s'agit là d'un principe de construction thématique qui, se dégageant au niveau des discours, dépend de la langue elle-même. De même, la notion de « stabilité », posée par les praticiens en termes de « continuité conceptuelle » et envisagée sur le seul plan lexical, peut être réinterprétée par l'examen des propriétés de langue en jeu dans la construction de la référence. De ce point de vue, la stabilité référentielle se comprend en termes d'effet d'interprétation : l'indexation apparaît alors comme une pratique destinée à construire de tels effets de stabilité référentielle.

Dès lors que le « monde » n'est plus pensé comme une donnée stable, l'indexation n'a plus pour fonction de « transmettre » une information qui lui préexisterait ; elle doit plutôt disposer d'éléments de nature à permettre la construction de l'information.

Par la confrontation systématique entre modes de représentation de la langue (des linguistes d'une part et des praticiens d'autre part), on parvient à distinguer les deux niveaux – du fonctionnement de la langue et de l'utilisation de la langue – en jeu dans l'indexation. Dès lors, on est en mesure de pouvoir spécifier ce qui caractérise en propre l'indexation. On peut rapporter aux propriétés de la langue elles-mêmes une partie de la pratique d'indexation ; on peut faire émerger parallèlement les zones où crucialement la pratique d'indexation se constitue comme telle.

Sur ces deux points, on peut montrer que :

- la viabilité de l'indexation repose sur l'exploitation des propriétés linguistiques suivantes : la signification lexicale, la synonymie référentielle, la construction de la référence discursive, la rigidité désignative. Cet ensemble de propriétés linguistiques fonde la possibilité de l'indexation, mais il ne garantit ni sa réalisation effective ni sa performance ;

- sur la base de ces propriétés de langue, l'indexation peut se donner un projet spécifique qui, du seul point vue linguistique, se laisse appréhender par la notion de construction de thèmes discursifs. La spécificité de l'indexation peut se définir en fonction d'un tel projet. Elle se fonde alors sur les caractéristiques de la construction des thèmes de discours : les thèmes de discours se construisent à travers plusieurs discours (notion d'interdiscours), et, plus précisément, à travers la saisie de plusieurs objets de discours (modèle des chaînes de référence). Le thème de discours relève en cela d'une lecture, d'une interprétation : il ne préexiste pas aux textes, dont il ne peut, en conséquence, être « extrait ». Le thème de discours se matérialise en outre sous une forme linguistique spécifique : le groupe nominal.

II - Spécificité de l'indexation : un espace de discours

Distingués des faits de langue, les faits d'indexation font apparaître leur singularité à deux niveaux (le document et le descripteur) qui relèvent tous deux d'une dimension discursive.

II.1 - Le discours documentaire

L'indexation apparaît d'abord comme un processus de création d'espaces documentaires spécifiques. Les questions qu'elle doit résoudre se dédoublent alors en deux problématiques :

- comment construire une collection documentaire qui permette une multiplicité de thématisations ?
- comme rendre accessible la collection documentaire ainsi constituée ?

Pour que ces deux questions puissent recevoir un cadre de réponse général dans lequel la particularité des pratiques puisse s'exprimer, on peut essayer de préciser le niveau de leur intervention, c'est-à-dire dégager un système de contraintes à partir duquel pourront être établies des règles spécifiques.

Concernant la constitution d'une collection documentaire, le principe directeur peut être défini à partir des conditions de l'existence et de la co-existence des textes eux-mêmes : à l'horizon théorique de l'indexation se place alors la notion d'« interdiscursivité ». L'examen des conditions qui déterminent l'existence des discours peut fournir des éléments de nature à guider la construction de l'interdiscours en indexation.

La problématique de l'accès à un espace spécifique de documents rejoint celle de la stabilité référentielle et peut être pensée à travers la réflexion menée en logique sur l'identité d'un objet. Cet arrière-plan théorique pose la question de la « reconnaissance » d'un « même » objet dans des termes sensiblement différents de ceux de l'appariement, de la rencontre entre expressions linguistiques semblables. La problématique de l'accès à une collection documentaire ne se réduit plus au choix du mot juste. L'approche logique montre en effet que l'identité d'un objet ne relève pas d'une mesure de ressemblance : elle suppose plutôt une construction qui permette à un objet de recevoir, tout en restant lui-même, de nouvelles propriétés. L'indexation apparaît alors comme cette construction susceptible de contraindre le parcours interprétatif des utilisateurs à travers l'espace documentaire qu'elle a constitué.

En tant qu'espace d'organisation spécifique des documents, l'indexation se laisse décrire comme un processus mettant en jeu deux types de stratégie : une stratégie d'exploration des sources et une stratégie d'exposition des documents. Ces deux stratégies s'établissent dans un objectif bien déterminé : la construction de thèmes de discours. Cet objectif relève moins de la recherche d'information proprement dite (le « contenu » des thèmes de discours construits par les utilisateurs n'intéresse qu'indirectement l'indexation) que d'un principe d'organisation des discours (sur quelles bases rapprocher des documents ? Les rendre complémentaires ? Les rendre accessibles ?).

On peut donc définir l'indexation non plus en fonction de son objectif (à quoi sert-elle ?) mais en fonction de ses objets : qu'est-ce que des textes, des documents ? Quels liens existent entre eux ? Comment permettre l'exploitation de ces liens ?

En prenant en compte la spécificité des objets qu'elle manipule, l'indexation peut trouver, dans ses propres objets, un « mode d'emploi » qui lui garantisse une certaine régularité de procédure. C'est la prise en compte de la condition d'existence des textes dans la stratégie d'exploration des sources ; c'est l'exploitation de la condition d'existence des objets de discours dans la stratégie d'exposition des documents.

II.2 - Le descripteur dans le discours documentaire

Dans le cadre de l'indexation comprise comme organisation spécifique de documents, la problématique du descripteur demande une nouvelle formulation. Le descripteur n'apparaît plus comme ce qui rapproche *a posteriori* des documents différents ; il se donne plutôt comme un élément qui permet de « circuler » dans un espace documentaire conçu *a priori* comme homogène. Cet élément doit présenter la caractéristique de pouvoir participer à la construction de thèmes discursifs. Cette contrainte détermine son rôle comme sa morphologie.

Le rôle du descripteur s'exprime clairement d'un point de vue logique : le descripteur doit établir une double relation de référence, et à une classe d'objets et à un objet particulier de cette classe. On peut évaluer sous cet angle le fonctionnement logique du nom propre et des descriptions définies. Cette évaluation engage à privilégier les descriptions définies de la logique qui demandent à être spécifiées d'un point de vue linguistique. L'étude des propriétés linguistiques qui doivent caractériser l'unité utilisée comme descripteur permet de capter une différence d'« état » entre le descripteur de l'indexation et le descripteur de la recherche documentaire. Si la forme linguistique du groupe nominal « complexe » constitue l'élément privilégié pour la construction du thème de discours en recherche d'information, c'est la forme linguistique de la synapsie qui permet aux indexeurs de pouvoir manipuler, organiser, trier, choisir les accès qu'ils mettent à disposition dans un système d'information.

Dans les deux cas, le descripteur se conçoit comme une unité extraite des textes eux-mêmes. C'est, là encore, en exploitant les discours eux-mêmes et la façon dont ils permettent la construction de la référence que l'indexation peut trouver « le mode d'emploi » du descripteur et les principes de sa définition. L'indexation, comprise sous le seul angle de la détermination des descripteurs, s'apparente alors à un processus d'extraction d'unités de discours qui idéalement se réalise de façon automatisée. Le projet de l'indexation – permettre la construction de thèmes de discours – présente cette caractéristique d'être inachevé d'un point de vue interprétatif. Capter des éléments susceptibles de participer à la construction d'un thème de discours suppose en effet un type de lecture qui se situe en deçà de l'interprétation : un indexeur humain peut difficilement se livrer à tel exercice.

C'est ainsi que la reformulation de l'indexation sur la base de fondements linguistiques rencontre une faisabilité technique actuelle : on dispose aujourd'hui de systèmes automatiques d'extraction d'unités de discours. Toutefois, de tels systèmes ne sauraient être considérés comme des systèmes d'indexation automatique. Tout au plus pourraient-ils servir dans le cadre d'une indexation assistée par ordinateur. L'indexation ne se définit pas uniquement sous l'angle de l'extraction d'unités de discours ; elle consiste essentiellement en l'organisation d'un espace de discours spécifique qui, lui, engage un travail scientifique sur les sources elles-mêmes, dont on envisage difficilement, pour le moment, les conditions d'automatisation.

II.3 - L'indexation dans le cadre des problématiques des « technologies de l'information »

Comme on le voit, l'indexation peut se laisser approcher sous un angle qui permet d'intégrer les « nouvelles technologies de l'information » sans s'y laisser ni réduire ni dissoudre. Il importe sur ce point d'insister sur ce qui constitue la spécificité de l'indexation, et plus généralement, celle des bibliothèques ou des centres de

documentation : « On veut parfois se passer de bibliothèque, de laboratoire, de collection sans perdre ni le savoir ni la raison. C'est croire à la "nature se dévoilant aux yeux de la science". Les chercheurs font bien autre chose que de contempler le monde dans un dérisoire *peep-show.* [...] C'est parce que les laboratoires, les bibliothèques et les collections se branchent sur un monde qui reste sans eux incompréhensible qu'il convient de les soutenir si l'on s'intéresse à la raison.[1] »

En mettant l'accent sur ce qui, en indexation, légitime le rapprochement de textes par des « mots » – le mode d'organisation des documents –, on se donne le moyen de pouvoir discuter la pertinence, en matière de recherche documentaire, des nouveaux réseaux de communication, comme le réseau Internet.

On peut en effet douter que le principe de fonctionnement du réseau Internet, tel qu'aujourd'hui il existe, soit en mesure d'améliorer, de faciliter la recherche d'information : à l'exigence documentaire d'organisation des sources s'oppose la volonté de « désenclaver » les données de leur source, de leur lieu de production, de leur temporalité. Si l'information documentaire doit trouver, sur les nouveaux réseaux de communication, une autre forme d'existence, il est nécessaire que les professionnels intègrent le débat, en mettant en avant les spécificités du traitement des documents qu'ils effectuent. Il importe sur ce point de pouvoir distinguer, dans l'indexation, une phase d'extraction des unités de discours et une phase, préalable et indispensable, d'organisation des discours.

III - De nouvelles pistes de recherche

Cette recherche a été conduite dans le but de faire émerger les propriétés linguistiques qu'exploitent les pratiques d'indexation, le plus souvent de façon implicite. Une fois circonscrits les aspects qui relèvent exclusivement du fonctionnement de la langue et du langage eux-mêmes se révèle l'étendue de ce qui constitue en propre l'indexation : ce que nous avons appelé sa stratégie d'exploration des sources et sa stratégie d'exposition des documents.

Pour spécifier ces deux aspects de l'indexation, l'approche linguistique ne suffit pas : d'autres disciplines doivent être convoquées.

Comment dégager des critères de sélection des sources qui tiennent compte de leurs conditions d'existence ? Le modèle des formations discursives proposé comme « horizon théorique » ne fournit que le principe général : prendre la mesure des « systèmes de la discursivité ».

Comment spécifier davantage de telles règles de sélection des sources ? Sans doute doit-on se tourner vers d'autres domaines de savoir, de type sociologique et historique, voire philosophique. On rencontre alors le paradoxe suivant que les fondements de l'indexation établis du point de vue de la théorie linguistique ne tiennent que sur la base du principe de l'interdiscours, mais cet interdiscours ne peut être spécifié du seul point de vue linguistique.

---

[1] Latour 1996, respectivement p. 43 et p. 45.

De même, le point de vue linguistique montre que la stratégie d'exposition des documents doit être pensée comme telle, en dehors des langages documentaires de type classificatoire qui reposent, tout comme l'indexation, sur des présupposés dont on peut montrer les limites. Vers quel autre type d'outil se tourner ? Les procédés d'exposition des connaissances dans le discours de vulgarisation scientifique fournissent quelques pistes : comment les exploiter dans le cadre d'un « discours » documentaire qui ne se « dit » pas mais qui fonctionne plutôt comme une transformation de discours existants ?

En outre, une réflexion menée d'un point de vue linguistique engage à distinguer, dans la notion d'accès documentaire, plusieurs niveaux : l'accès à un ensemble de documents et l'accès à l'« information » ne mettent pas en jeu les mêmes procédés. On a proposé que l'accès à un ensemble de documents se fasse d'abord par le biais de « domaines ». L'accès à l'information s'établit, lui, sur la base des éléments d'une chaîne de référence. Comment articuler ces différents types d'accès ?

Le seul point de vue linguistique, s'il permet d'aboutir à de telles conclusions, ne permet pas toujours de les spécifier.

Au terme de cette recherche apparaissent une multitude de questions, laissées ici en suspens. Ces questions nécessitent d'être résolues pour que l'indexation, explicitement fondée d'un point de vue linguistique, puisse pouvoir être effectivement pratiquée. La seule mise en valeur des propriétés linguistiques à l'œuvre en indexation ne suffit pas toujours à modifier les pratiques existantes : extraire, des textes, des syntagmes nominaux ou des synapsies ne permettra pas, nécessairement, d'améliorer de façon significative l'indexation et la recherche documentaires. Il s'agit là d'une condition nécessaire mais pas suffisante.

En outre, pour pouvoir maîtriser les deux types de stratégie par lesquels nous avons proposé de concevoir l'indexation, il est indispensable de pouvoir définir de façon précise les notions de « source documentaire » et de « document » : nous avons indiqué quelques propriétés susceptibles de permettre de distinguer les objets « indexables » d'une part et de caractériser les objets indexés d'autre part. Mais là encore, il ne s'agit que d'une amorce : beaucoup reste à faire.

L'ampleur de la tâche à accomplir peut néanmoins être réduite par la prise en compte des travaux menés sur d'autres pratiques professionnelles. Sur ce point, la pratique de la terminologie et celle de la vulgarisation nous sont apparues de nature à pouvoir être profitablement rapprochées des problématiques documentaires. Cependant, là encore, il est nécessaire de pouvoir préciser le « continuum de la diffusion des connaissances » qui traverserait terminologie, indexation et vulgarisation scientifique : où placer l'indexation ? Comment se singularise son mode de diffusion de la connaissance ?

Les questions à résoudre qui permettraient à nos propositions de trouver une voie de réalisation effective sont nombreuses et importantes. Reste que l'approche de l'indexation proposée dans cette étude montre qu'il est possible de spécifier l'indexation dans le rapport particulier qu'elle entretient avec les textes eux-mêmes et les unités lexicales qui les constituent : cette tentative de « réconciliation » entre la langue et la pratique professionnelle de l'indexation nous semble en cela constituer une hypothèse de travail constructive.

Annexes

ANNEXE 1 :
PRÉSENTATION DE L'EXPÉRIMENTATION

1 - Objectif de l'expérimentation

L'expérimentation menée avait pour objectif d'apporter des éléments de réponse aux deux questions suivantes :

- une même source peut-elle faire l'objet de « mises en document » différentes ?
- s'il y a une diversité de mises en documents, à quel niveau et sous quelle forme se manifeste-t-elle ?

Pour appréhender ces deux questions, nous avons procédé à l'analyse comparative du traitement documentaire d'une même source.

Par traitement documentaire, on entend :

- la sélection des unités textuelles à indexer ;
- l'affectation de descripteurs aux unités textuelles retenues ;
- l'objectif d'utilisation assigné aux unités textuelles sélectionnées et indexées.

2 - Méthode

2.1 - Choix de la source

Pour observer le mécanisme de sélection documentaire, nous avons choisi comme source un périodique généraliste, supposé dense et varié : le quotidien *Le Monde*. L'édition retenue a été celle du 1er décembre 1994, date qui convenait aux

participants sollicités et qui présentait l'avantage de comporter un supplément, lui toujours spécialisé[1].

Ne nous étant pas livrée à une analyse fine des publications du *Monde*, il est difficile d'évaluer la représentativité de l'édition retenue et par suite de l'expérimentation menée.

À titre indicatif, signalons quelques caractéristiques du numéro traité :
- numéro de 24 pages dont deux pages « non informatives » (publicité, p. 5 ; annonces immobilières p. 17) ;
- numéro comportant 92 « textes » dont :
  - 53 articles signés
  - 7 encadrés de rappel[2]
  - 16 dépêches identifiables (AFP ou Reuter)
  - 16 articles non signés
  - 5 hors-textes
  - 1 rectificatif

2.2 - Choix des participants

Le type de source retenu a orienté le choix des participants : ce sont essentiellement les centres de documentation qui se livrent au dépouillement de la presse quotidienne.

Le nombre de participants a arbitrairement été arrêté à dix, une fois obtenue la diversité souhaitée :
- diversité de statut : 6 organismes relèvent du secteur privé, 4 du secteur public ;
- diversité des couvertures documentaires : 3 organismes documentaires généralistes, 7 spécialisés ;
- diversité des langages documentaires : 2 thésaurus, 1 liste de vedettes matières, 2 plans de classification, 6 lexiques de classement ;
- diversité de l'utilisation des articles du Monde : alimentation de bases de données bibliographiques, alimentation de dossiers de presse, fabrication de revues de presse, diffusion sélective d'information[3].

La liste des dix participants est la suivante :

	Statut	**Couverture documentaire**	**Langage documentaire**	**Utilisation des articles**
Capital	Privé Presse	Micro-économie Entreprises	Lexique de classement	Alimentation dossiers / DSI
CNDP Centre national de documentation pédagogique	Public Éducation nationale	Communication	Lexique de classement	Revue de presse bimensuelle

[1] Le supplément du jour concernait les arts et spectacles (ancienne formule du journal).
[2] Nous avons distingué article et encadré de rappel car les documentalistes n'ont pas toujours sélectionné les deux.
[3] Abrégée DSI dans le tableau ci-dessous.

Cour des Comptes	Public Juridiction financière	Financements publics	Plan de classification	Revue de presse quotidienne Alimentation dossiers
La Croix	Privé Presse	Généraliste	Lexique de classement	Revue de presse quotidienne Alimentation dossiers
Documentation française	Public Service du Premier ministre	Information politique Actualité	Thésaurus	Alimentation de base de données
Le Figaro (documentation économique)	Privé Presse	Économie Entreprises	Lexique de classement	Alimentation dossiers / DSI Revue de presse quotidienne
FNSP Fondation nationale des sciences politiques	Public Enseignement supérieur	Sciences politiques Actualité	Plan de classification Liste de vedettes matières (Rameau)	Alimentation dossiers Alimentation de catalogue
InfoMatin	Privé Presse	Généraliste	Lexique de classement	Alimentation dossiers
Le Monde	Privé Presse	Généraliste	Thésaurus	Archivage du journal Alimentation dossiers
Télérama	Privé Presse	Culture Médias Société	Lexique de classement	Alimentation dossiers Revue de presse quotidienne

3 - Déroulement de l'expérimentation

3.1 - Calendrier

L'expérimentation s'est déroulée approximativement d'octobre 1994 à octobre 1995 :

Septembre 1994	*Contacts et accord des participants*
Octobre 1994	*Mise au point des consignes et du questionnaire*
Novembre 1994	*Envoi des consignes et du questionnaire*
Mars 1995	*Recueil des réponses et entretiens*
Août-septembre 1995	*Dépouillement et analyse des résultats*

3.2 - Consignes et questionnaire

Les consignes de travail étaient les suivantes :
- reporter le travail normalement effectué dans le cadre du service sur l'exemplaire du *Monde* fourni, en indiquant : le cochage, l'indexation et l'utilisation envisagée ;
- indiquer le temps réel passé au dépouillement du journal ;
- suivre le traitement documentaire du numéro : indiquer les modifications apportées à la sélection et/ou à l'indexation et/ou à l'utilisation au cours des quatre mois suivants.

Le questionnaire joint aux consignes portait sur :
- la couverture documentaire ;
- les périodiques dépouillés au même titre que *Le Monde* ;
- le mode de stockage du journal (texte intégral, coupures de presse ; supports imprimé, électronique, magnétique) ;
- outils et méthodes documentaires employés ;
- une rubrique de commentaire libre sur le dépouillement réalisé.

ANNEXE 2 :
LES MISES EN DOCUMENTS

Cette annexe présente, sous une forme chiffrée, le résultat issu de l'analyse du cochage effectué ; elle met en regard le nombre d'articles retenus (colonne « Nbr. art. ») et le nombre d'unités documentaires constituées à partir de ceux-ci (colonne « Nbr. doc. »).

Nom de l'organisme	**Mise en document**		**Observations**		
	Nbr. art.	Nbr. doc.	Temps passé	Stockage du Monde	Autres
Capital	5	3	10 mn	Texte intégral imprimé (1 an) Coupures	Le Monde par Minitel
CNDP	2	1	10 mn	Texte intégral imprimé (8 jours)	
Cour des comptes	6	6	15 mn	Texte intégral imprimé ("longtemps") Coupures CD-Rom	Journal de référence
La Croix	41	42	2 heures	Texte intégral imprimé ("longtemps") Coupures	Journal de référence
Documentation française	22	13	1h30	Coupures de presse (microfiches)	Journal de référence
Le Figaro-Eco	6	4	15 mn	Texte intégral imprimé (1 an) Coupures	
FNSP	78	85	4 heures	Texte intégral imprimé (8 jours) Coupures	Journal de référence Pôle associé Sciences Politiques
Info Matin	44	37	45 mn	Texte intégral imprimé (1 an)	Le Monde par l'Européenne de Données Abont. AFP
Le Monde	89	89			
Télérama	30	28	30 mn	Texte intégral imprimé (6 mois)	Le Monde par l'Européenne de Données

ANNEXE 3 :
LES NOMS PROPRES
DANS LES PRATIQUES DOCUMENTAIRES

La place des noms propres dans les pratiques documentaires peut être appréhendée à trois niveaux :

- au niveau global des grilles d'indexation utilisées ;
- au niveau médian de la distribution des termes d'indexation ;
- au niveau local de la nature linguistique des descripteurs matières.

1 - Présence de la catégorie « nom propre » dans les grilles d'indexation

Organisme documentaire	Grille d'indexation
La Croix	Comprend deux champs d'indexation réservés aux noms propres : - lieu géographique de l'action - personnes citées
Le Monde	Constituée de quatre champs d'indexation imposant l'emploi de noms propres : - territoires étrangers - territoires français - personnes physiques, personnes morales - auteurs dont l'œuvre est citée
Documentation française	Comprend un champ obligatoire pour les descripteurs géographiques
FNSP	Le plan de classification impose l'entrée par un nom géographique[1]
Cour des comptes	Adopte le plan de classification de la FNSP

[1] « Chaque fois que cela est possible on classe d'abord sous le nom d'un pays, ou sous celui d'un ensemble géographique et seulement si aucune localisation n'est possible dans les rubriques générales (9.1 et 9.3 à 9.7). Dans ces cas, le chiffre 9 prend la place du nom de pays ».

2 - Distribution des termes d'indexation

Abréviations utilisées dans le tableau :

- Nbre de I : rappel du nombre global d'indexations réalisées ;
- Nbre de G : indication du nombre de noms de lieux géographiques (pays, villes, régions, etc.) utilisés pour indexer ;
- Nbre de PP : indication du nombre de noms de personnes physiques utilisés pour indexer ;
- Nbre de PM : indication du nombre de noms de personnes morales (entreprises, manifestations culturelles, institutions, partis politiques, etc.) utilisés pour indexer ;
- Nbre de M : indication du nombre de « descripteurs matières » utilisés pour indexer ;
- % : part des noms propres dans les indexations.

Tableau de distribution

Organisme	Nbr I	NbrG	Nbr PP	Nbr PM	NbrM	%
Capital	3		2	1		100 %
CNDP	1			1		100 %
Cour des comptes	6					
La Croix	42	14	9	3	16	62 %
Doc. française	13					50 %
Le Figaro-Éco	4	1		2		100 %
FNSP	85					94 %
Info Matin	37	15	8	2	12	67,5 %
Le Monde	89				11	88 %
Télérama	28	7	5	3	5	75 %

Commentaire du tableau

La distribution proposée ci-dessus est incomplète et donc approximative : la grille d'analyse retenue ne permet pas en effet de rendre compte complètement de tous les cas de pratiques documentaires ; certaines évoluent dans le temps, d'autres sont trop complexes pour être rendues ici dans le détail :

- *exemple de pratique complexe* : La Documentation française dispose de règles de composition de descripteurs fondées sur le mélange nom propre-nom commun[1] (exemple : « PS – parti politique position ») ; elle travaille en outre sur deux niveaux d'analyse (détaillé et général) partiellement redondant : nous avons essayé d'estimer plus que de comptabiliser la proportion des noms propres ;

- *exemple d'évolution de l'indexation dans le temps* : élimination d'articles une fois diffusés par revue de presse (cas de Télérama, de la Cour des comptes).

[1] On distingue ces règles de composition « nom commun-nom propre » des entrées de lexiques de classement constituées, de façon figée, d'un nom propre et d'un nom commun.

3 - Nature linguistique des descripteurs thématiques

En examinant la composition des descripteurs thématiques (colonne « nbr M » dans le tableau ci-dessus), on peut établir une distinction entre les descripteurs thématiques comportant uniquement des « noms communs » et les descripteurs thématiques comportant à la fois des « noms communs » et des « noms propres ».

Abréviations utilisées dans le tableau

- Nbre M : indication du nombre d'indexations thématiques ;
- M NC : descripteurs thématiques constitués uniquement de noms communs ;
- M NC-NP : descripteurs thématiques constitués à la fois de noms communs et de noms propres.

Nom	Nbr M	M NC	M NC et NP
La Croix	16[1]	- institutions - corruption - sida / conférence - avortement - enseignement supérieur - chômage / piste	- organisation internationale / OCDE - élections présidentielles 1995 / Droite / Séguin - ONU - organisation internationale / OTAN - drogue / France - cinéma / festival / Nantes - organisation internationale / OMC
Info Matin	12	- élections présidentielles / primaires - politique / financement - drogue / dépénalisation - IVG / commandos - immigration - cinéma / festivals - universités	- télévision / Arte - presse / Actuel - organisation internationale / OCDE - GATT - transport maritime / Achille Lauro
Le Monde	11	- analyse, chiffre ; emploi transitoire ; jeunesse / chômage, octobre, indice mensuel, 1994.	- Bosnie, affrontement ethnique, États-Unis, comportement, politique étrangère, déclaration. - Russie, prise de position politique, constitution, souveraineté. - Koweït, adoption parlementaire, projet de loi, femme, université, interdiction. - Grande-Bretagne, projet, budget, 1995, 1996. - sommet politique du gouvernement, dialogue nord-sud, pays en voie de

[1] Trois de ces seize indexations sont identiques.

			développement, signature, projet, Afrique, Russie. - Venezuela, décès, prison, évasion, chiffre. - Éthiopie, adoption parlementaire, État fédéraliste, liste. - Burundi, attentat, décès - Nigeria, décès, date, élection législative, 1995. - Zimbabwe, accord, force d'interposition, Angola.
Télérama	5	- institution - drogue - sida - enseignement - organisation internationale	

GLOSSAIRE

La première occurrence, dans notre texte, des termes ci-dessous est marquée d'une étoile (*).

Analyse terminologique

Identification des notions appartenant à un domaine donné, et étude en contexte des termes qui les désignent et des relations qui les sous-tendent.

Se déroule en quatre étapes :

1 - Découpage terminologique : identification du statut terminologique d'une unité extraite d'un énoncé ;

2 - Analyse contextuelle : délimitation du contenu notionnel d'un terme en contexte par l'identification et l'analyse des caractères de la notion présents dans ce contexte ;

3 - Recoupement notionnel : conformité des caractères notionnels contenus dans les définitions et généralement dans les contextes relevés, qui permet d'établir la correspondance entre les notions et l'équivalence ou la synonymie entre les termes ;

4 - Analyse notionnelle : détermination des caractères d'une notion, de sa compréhension, de son extension et des relations qu'elle entretient avec d'autres notions.

[Boutin-Quesnel et *al.* 1990]

Anaphore

Un segment de discours est dit anaphorique lorsqu'il fait allusion à un autre segment, bien déterminé, du même discours, sans lequel on ne saurait lui donner une interprétation (même simplement littérale). [Ducrot et Schaeffer 1995]

Approche classique de l'indexation

Discours de la pratique sur elle-même.

Appelé aussi « discours classique ».

Concept en terminologie

On peut dire que le concept général n'est ni un simple signe ni une idée véritable [...] mais qu'il consiste en un schème opératoire de notre entendement. [Rey 1992]

Contexte

I - Entourage linguistique d'un élément (d'une unité phonique dans un mot, d'un mot dans une phrase, d'une phrase dans un texte) ; appelé aussi « cotexte ».

II - Ensemble des circonstances au milieu desquelles a lieu une énonciation (écrite ou orale) ; appelé aussi « situation de discours » ou encore « contexte situationnel ».

[Ducrot et Schaeffer 1995]

Dénomination

Attribution d'un signe à un élément du réel pour pouvoir à l'aide de ce signe y référer ensuite durablement, étant entendu qu'un signe ayant un statut dénominatif appartient nécessairement à la catégorie nominale. [Corbin 1993]

Dénotation

Ensemble des référents d'un signe (*Bedeutung*). [Frege 1971 [1892]]

Descripteur

Mot ou groupe de mots retenus dans un thésaurus et choisis parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande d'information. [AFNOR 1978]

Discours documentaire

Espace d'organisation des documents établi sur un *a priori* non formel.

Document

I - Selon l'AFNOR : Ensemble d'un support d'information, des données enregistrées sur ce support et de leur signification, servant à la consultation, l'étude, la preuve. [AFNOR 1978]

II - Selon Escarpit : Signe, ensemble de signes ou message fixés au moyen de traces sur un support ; s'oppose à « semi-document »*. [Escarpit 1991]

Énoncé

Segment d'un discours ou discours en dehors de ses conditions de production. [Souchard 1989]

Énonciation

Production d'un segment d'un discours ou discours et les conditions de cette production. [Souchard 1989]

Extension en logique

Classe des objets dénotés par un signe. [Mounin 1974]

Intension en logique

Ensemble des traits qui définissent la classe dénotée par le signe. [Mounin 1974]

Interprétation d'un texte

I - Selon la sémantique interprétative : assigner un sens aux textes à partir de leur matérialité.

II - Selon la philosophie : herméneutique.

[Rastier 1994]

Langage

I - S'il est entendu en tant que « factum loquendi » (il y a des êtres parlants), le langage ne constitue pas un objet pour la science linguistique ; il est celui de la philosophie du langage ;

II - S'il est entendu comme le support des propriétés de langue, le langage constitue l'objet de la science linguistique.

[Milner 1989]

Langage documentaire

Langage artificiel constitué de représentations de notions et de relations entre ces notions et destiné, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs. [*Vocabulaire de la documentation* 1987]

Langue

Le terme *langue* « sténographie » un complexe de trois faits :

I - le « factum linguae » : le fait que ce que parle un être parlant mérite le nom de langue ;

II - le « factum linguarum » : le fait que les langues soient diverses, tout en constituant une classe homogène ;

III - le « factum grammaticae » : le fait que les langues soient descriptibles en termes de propriétés.

[Milner 1989]

Linguistique

Ensemble de postulats sur la langue (ou le langage) et de méthodologies de description. [Marandin 1979]

Mot construit

Mot dont le sens prédictible est entièrement compositionnel par rapport à la structure interne, et qui relève de l'application à une catégorie lexicale majeure (base) d'une opération dérivationnelle [...] associant des opérations catégorielle, sémantico-syntaxique et morphologique. [Corbin 1987]

Objet de discours

Notion empruntée à la logique naturelle de Grize.

Objets produits par l'activité cognitive et interactive des sujets parlants.

[Apothéloz et Reichler-Béguelin 1995]

Objet scientifique

Objet qui résulte d'une schématisation de l'expérience et de son insertion dans un système de concepts où il prend sens, et qui lui sert en quelque sorte de référentiel. [Granger 1993]

Recherche documentaire

Pour les uns, la recherche documentaire se limite à l'opération par laquelle les documents sont choisis dans le fonds documentaire à la demande de l'usager ; pour d'autres, elle consiste à fournir, en fonction d'une demande définie et spécifique formulée par l'usager, les éléments d'information documentaire correspondants. Pour certains, enfin, la recherche documentaire est une réponse plus ou moins élaborée à une demande et elle doit aboutir à un produit dont la forme est décidée avec l'usager (bibliographie, note de synthèse, etc.). [Sutter 1997a]

Référence

Propriété d'un signe linguistique lui permettant de renvoyer à un objet du monde extra-linguistique, réel ou imaginaire. [*Dictionnaire de linguistique et des sciences du langage* 1994]

Référence actuelle

Relation entre une molécule syntaxique [un groupe nominal] et un objet possible du monde. [Milner 1989]

Référence virtuelle

Ensemble de conditions que doit satisfaire un objet du monde pour pouvoir être désigné, en référence actuelle, par une molécule syntaxique dont l'atome syntaxique sera le Nom principal. [Milner 1989]

Représentation [collective]

Conceptions et symboles qui résultent de l'interaction sociale et qui acquièrent une signification commune pour les membres du groupe en provoquant chez eux des réactions émotionnelles semblables. [Willems 1970]

Sciences du langage

Voir *Linguistique*

Semi-document

Document qui maintient la diachronie interne de l'événement qu'il stabilise dans le temps. [Escarpit 1991]

Sens

I - C'est le mode de donation de l'objet (*Sinn*), s'oppose à dénotation : ensemble des référents d'un signe (*Bedeutung*). [Frege 1971 [1892]]

II - La notion de *Sinn* concerne la manière dont l'être vrai d'une expression peut être déterminé à partir de sa constitution interne. Le *Sinn* d'une expression X donnée décrit la manière dont, en tant que partie d'une expression complète Y, X contribue à l'être vrai de Y. [Milner 1989]

Signification lexicale

Propriété distinctive attribuable à un atome syntaxique.

Référence virtuelle* d'un atome syntaxique.

[Milner 1989]

Source

[Terme pouvant être rapproché de « document primaire »]

Document qui présente une information à caractère original, c'est-à-dire lue ou vue par le lecteur dans le même état où l'auteur l'a écrite ou conçue. [Sutter 1997b]

Taux de précision

Proportion de documents trouvés qui sont pertinents par rapport au total des documents trouvés. [Kent, Berry et *al.* 1955]

Taux de rappel

Proportion de documents pertinents qui sont trouvés par rapport au total des documents pertinents de la collection. [Kent, Berry et *al.* 1955]

Terme en terminologie

Unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe) et qui désigne une notion de façon univoque à l'intérieur d'un domaine. [Boutin-Quesnel et *al.* 1990]

Théorie

Le but des théories est d'exprimer des régularités et, de façon générale, d'apporter une compréhension plus approfondie des phénomènes en question. On fait ensuite la supposition que ces derniers sont régis par des lois théoriques ou par des principes théoriques caractéristiques, grâce auxquels la théorie explique alors les relations uniformes antérieurement découvertes et prédit aussi des régularités « nouvelles » du même ordre. [Hempel 1996 [1972]]

Théorie du lexique

Théorie des unités lexicales hors emploi. [Marandin 1992a]

Thésaurus

Liste d'autorité organisée de descripteurs et de non-descripteurs obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques, hiérarchiques, associatives, ou d'équivalence. Cette liste sert à traduire en un langage artificiel dépourvu d'ambiguïté des notions exprimées en langage naturel. [AFNOR 1981]

Unité lexicale (ou terme, dans la théorie de la syntaxe positionnelle)

Ensemble de toutes les propriétés distinctives d'un atome syntaxique : appartenance catégorielle ; forme phonologique ; signification lexicale*. [Milner 1989]

BIBLIOGRAPHIE

AFNOR 1978. « Norme Z 47-102 : principes généraux pour l'indexation des documents » *in Documentation.* Tome 1 : présentation des publications, traitement documentaire et gestion des bibliothèques. Paris : AFNOR. (Recueil de normes françaises).

AFNOR 1981. « Norme Z 47-100 : règles d'établissement des thésaurus monolingues » *in Documentation.* Tome 1 : présentation des publications, traitement documentaire et gestion des bibliothèques. Paris : AFNOR. (Recueil de normes françaises).

AFNOR 1985. « Norme Z 47-200 : liste d'autorité de matières » *in Documentation.* Tome 1 : présentation des publications, traitement documentaire et gestion des bibliothèques. Paris : AFNOR. (Recueil de normes françaises)

AFNOR 1995. « Norme AFNOR Z 44-077 : catalogage des images » *in Documentation.* Tome 3 : catalogage des non-livres. Paris : AFNOR. (Recueil de normes françaises).

APOTHÉLOZ (Denis) et Marie-José Reichler-Béguelin, 1995. « Construction de la référence et stratégies de désignation ». *TRANEL*, n° 23, p. 227-271.

AUPELF-UREF 1994. *Ingénierie de la langue* [texte de l'appel d'offres, n.p.].

AUROUX (Sylvain) 1994. « Encyclopédies, bibliothèques, formalisation du savoir » *in Science en bibliothèque* (sous la dir. de Francis Agostini). Paris : Éditions du Cercle de la librairie. (Collection Bibliothèques). P. 141-150.

BATIME (Christine) 1995. « Le langage vu comme un " référentiel de représentations négociées " constitutif d'un système d'information au sein d'une organisation ». *Cahiers de linguistique sociale.* (Actes du colloque « Recherches documentaires »), p. 19-26.

BAUDELOT (Christian) et Claire Véry 1994. « Profession : lecteur ? Résultats d'une enquête sur les lecteurs de la Bibliothèque nationale ». *Bulletin des bibliothèques de France*, n° 4, p. 8-17.

BEAULIEU (Françoise) 1995. « Dis-moi ton nom, je te dirai qui tu es : réflexion sur les pouvoirs de dénomination ». *Recherches sémiotiques*, vol. 15, n° 12, p. 35-48.

BEGHTOL (Clare) 1986. « Bibliographic classification theory and text linguistics : aboutness analysis, intertextuality and the cognitive act of classifying documents ». *Journal of Documentation*, vol. 42, n° 2, p. 84-113.

BELY (N.), A. Borillo, S. Siot-Decauville et J. Virbel, 1970. *Procédures d'analyse sémantique appliquées à la documentation scientifique.* Paris : Gauthier-Villars.

BENOÎT (Denis) sous la dir. de, 1992. *Introduction aux sciences de l'information et de la communication.* Paris : Éditions d'Organisation.

BENVENISTE (Émile) 1974 [1969]. « Sémiologie de la langue » *in Problèmes de linguistique générale.* Tome 2. Paris : Gallimard. (Tel). P. 43-66.

BENVENISTE (Émile) 1974 [1966]. « Formes nouvelles de la composition nominale » *in Problèmes de linguistique générale.* Tome 2. Paris : Gallimard. (Tel). P. 163-176.

BERRENDONNER (Alain) 1978. *Les référents nominaux du français et la structure de l'énoncé.* Thèse de doctorat. Lyon : Université de Lyon II.

BERRENDONNER (Alain) 1981. *Éléments de pragmatique linguistique.* Paris : Éditions de Minuit.

BERRENDONNER (Alain) 1983. *Grammaire pour un analyseur : aspects morphologiques.* Document de travail du groupe SYDO [n.p.].

BERRENDONNER (Alain) 1995a. « Quelques notions utiles à la sémantique des descripteurs nominaux ». *TRANEL*, n° 23, p. 9-39.

BERRENDONNER (Alain) 1995b. « Redoublement actanciel et nominalisations ». *SCOLIA*, n° 5, p. 215-244.

BERRENDONNER (Alain), Richard Bouché, Michel Le Guern et Jacques Rouault, 1980. « Pour une méthode d'interaction pondérée des composants morphologique et syntaxique en analyse automatique du français ». *T.A. Informations*, n° 1, p. 3-28.

BERRENDONNER (Alain), Michel Le Guern et G. Puech, 1983. *Principes de grammaire polylectale.* Lyon : Presses universitaires de Lyon.

BERRENDONNER (Alain) et Marie-José Reichler-Béguelin, 1989. « Décalages : les niveaux de l'analyse linguistique ». *Langue française*, n° 81, p. 99-125.

BERTRAND (Annick) 1993. *Compréhension et catégorisation dans une activité complexe : l'indexation des documents scientifiques.* Thèse de doctorat. Toulouse : Université de Toulouse-Mirail.

BERTRAND-GASTALDI (Suzanne) 1986. « De quelques éléments à considérer avant de choisir un niveau d'analyse ou de langage documentaire ». *Documentation et bibliothèque*, vol. 32, n° 1-2, p. 3-23.

BERTRAND-GASTALDI (Suzanne) 1989. « La Problématique de l'énonciation dans les systèmes documentaires entièrement ou partiellement automatisés ». *Cahiers Recherches et Théories* (numéro consacré aux « Problèmes de l'énonciation », sous la dir. de François La Traverse), p. 9-80.

BERTRAND-GASTALDI (Suzanne) 1993. « Analyse documentaire des jugements et intertextualité », *in Les Sciences du texte juridique : le droit saisi par ordinateur* (Séminaire Val Morin, 5-7 octobre 1992). Québec : Y. Blais. P. 139-173.

BIBLIOTHÈQUE NATIONALE DE FRANCE Service de la coordination bibliographique, 1995. *Guide d'indexation RAMEAU.* Montpellier : ABES.

BLAIR (D.R.) 1990. *Language and Representation in Information Retrieval.* Netherlands : Elsivier science publishers.

BLANQUET (Marie-France) 1994. *Intelligence artificielle et système d'information.* Paris : ESF. (Systèmes d'information et nouvelles technologies).

BOLTANSKI (L.) et P. Maldidier 1977. *La Vulgarisation scientifique et son public, une enquête sur Sciences et Vie.* Paris : Éditions de l'École des Hautes Études en Sciences Sociales.

BONHOMME (Marc) 1987. *Linguistique de la métonymie.* Berne : Peter Lang. (Sciences pour la communication ; 16).

BOUCHÉ (Richard) 1988. « Sciences de l'information : sciences de la mise en forme » *in Infomédiatique.* Paris : Éditions du Cercle de la librairie. P. 11-18.

BOUCHÉ (Richard) 1989. « Le syntagme nominal, une nouvelle approche des bases de données textuelles ». *Méta,* vol XXXIV, n° 3, p. 428-434.

BOUDON (Raymond) 1986. *L'Idéologie ou l'origine des idées reçues.* Paris : Fayard.

BOUGNOUX (Daniel) 1993. *Sciences de l'information et de la communication.* Paris : Larousse. (Textes essentiels).

BOURDIEU (Pierre) 1982. *Ce que parler veut dire : l'économie des échanges linguistiques.* Paris : Fayard.

BOURION (Évelyne) et Denise Malrieu, 1994. « Concepts, systèmes signifiants et organisation d'un domaine : étude sémantique et sémiotique d'un plan de classement de base de données ». *Cahiers de lexicologie,* vol. 1, n° 64, p. 83-131.

BOUTAYEB (Samy) 1993. « Terminologie et documentation : au confluent de deux disciplines » *in* Séminaire *Terminologie et documentation,* 7-9 juin 1993. Paris : Institut d'Études Politiques de Paris. [Support de cours ; n.p.].

BOUTIN-QUESNEL (Rachel), Nicole Bélanger, Nada Kerpan, Louis-Jean Rousseau, 1990. *Vocabulaire systématique de la terminologie.* Québec : Les Publications du Gouvernement du Québec . (Cahiers de l'Office de la langue française).

BRUXELLES (Sylvie) 1991. « Codage et construction du sens : approche linguistique des processus interprétatifs mis en œuvre dans l'application d'une nomenclature des contentieux judiciaires civils » *in Sémantique et cognition : catégories, prototypes, typicalité* (sous la dir. de Danièle Dubois). Paris : Éditions du CNRS. (Sciences du langage). P. 171-186.

BUISSON (Corinne) 1995. « L'indexation : de la technique à la pratique ». *Cahiers de linguistique sociale.* (Actes du colloque « Recherches documentaires »), p. 121-128.

CALENGE (Bertrand) 1994. *Les Politiques d'acquisition : constituer une collection dans une bibliothèque.* Paris : Éditions du Cercle de la librairie. (Collection Bibliothèques).

CAPURRO (Rafael) 1992. « What is information science for ? A philosophical reflexion » *in Conception of library and information science : historical, empirical and theorical perspectives.* London : Taylor Graham. P. 84-102.

CARRÉ (René), Jean-François Degremart, Maurice Gross, Jean-Marie Pierrel et Gérard Sabah, 1991. *Langage humain et machine.* Paris : Presses du C.N.R.S.

CAUQUELIN (Anne) 1990. *Aristote et le langage.* Paris : Presses universitaires de France. (Philosophies).

CHASTAIN (Charles) 1975. « Reference and context ». *Language Mind and Context.* [s.l.] : K. Gunderson. P. 194-273.

CHAUDIRON (Stéphane) 1994. « L'intégration des technologies d'ingénierie linguistique dans le traitement de l'information » in *Congrès IDT,* 31 mai-2 juin 1994. Paris : ADBS ; ANRT. P. 100-104.

CHAUMIER (Jacques) 1978. *Les Langages documentaires : le traitement linguistique de l'information documentaire.* Paris : Entreprise moderne édition.

CHAUMIER (Jacques) 1988. *Le Traitement linguistique de l'information.* Paris : Entreprise moderne édition.

CHAUMIER (Jacques) 1989. *Les Techniques documentaires.* Paris : Presses universitaires de France. (Que sais-je ? ; 1419).

CHAUMIER (Jacques) 1996. *Travail et méthodes du documentaliste.* Paris : ESF.

CHOMSKY (Noam) 1981 [1975]. *Réflexions sur le langage.* Paris : Flammarion. (Champs ; 46).

CLEVERDON (Cyril), Jack Mills and Michael Keen, 1966. *Cranfield Research Project.* [s.l.] : National Science Foundation.

COLLARD (Claude), Isabelle Gianattasio et Michel Melot, 1995. *Les Images dans les bibliothèques.* Paris : Éditions du Cercle de la librairie. (Collection Bibliothèques).

COLLAS (Dominique) et Ghislaine Chartron, 1994. « Logique conceptuelle et recherche d'information ». *Documentaliste-Sciences de l'information*, vol. 31, n° 1, p. 9-15.

COMARONI (John) 1988. *Dewey Decimal Classification : history and current status.* [s.n.] : Envoy Press.

COMITÉ NATIONAL D'ÉVALUATION (CNE) 1993. *Les Sciences de l'information et de la communication.* Paris : CNE.

CORBIN (Danielle) 1987. *Morphologie dérivationnelle et structuration du lexique.* Tübingen : Max Niemeyer.

CORBIN (Danielle) 1989. « Contraintes et création lexicales en français ». *L'Information grammaticale*, n° 42, p. 35-43.

CORBIN (Danielle) 1990. « Homonymie structurelle et définition des mots construits : vers un dictionnaire "dérivationnel" » *in La Définition* (actes du colloque CELEX). Paris : Larousse. P. 175-192.

CORBIN (Danielle), Georgette Dal, Agnès Mélis-Puchulu et Martine Temple, 1993. « D'où viennent les sens a priori figurés des mots construits ? Variations sur lunette(s), ébéniste et les adjectifs en -esque ». *Verbum*, n° 1-2-3, p. 65-100.

CORBIN (Danielle) et Martine Temple, 1994. « Le Monde des mots construits et des sens construits : catégories sémantiques, catégories référentielles ». *Cahiers de lexicologie*, vol. 65, n° 2, p. 5-28.

CORBLIN (Francis) 1987. « Les Chaînes de référence naturelles ». *T.A. Informations*, n° 1, p. 5-21.

CORBLIN (Francis) 1995. *Les Formes de reprise dans le discours : anaphores et chaînes de référence.* Rennes : Presses universitaires de Rennes. (Collection Langue/discours).

CORI (Marcel) et Jean-Marie Marandin, 1993. « Grammaire d'arbres polychromes ». *T.A.L.*, n° 1, p. 101-132.

COYAUD (Maurice) 1966. *Introduction à l'étude des langages documentaires.* Paris : Klincksieck.

COYAUD (Maurice) et N. Siot-Decauville, 1972. *Linguistique et documentation. Les articulations logiques du discours.* Paris : Larousse.

CROS (R.-C.), J.-C. Gardin et F. Lévy, 1964. *L'Automatisation des recherches documentaires : un modèle général, le SYNTHOL.* Paris : Gauthier-Villars.

DACHELET (Roland) 1990. *État de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information.* Le Chesnay : INRIA. (Rapport de recherche ; 1201).

DAVID (Sophie) 1989. *Évaluation de la thèse de C.L. Sidner et propositions pour un traitement automatique des anaphores.* DEA de linguistique théorique et formelle. Paris : Université de Paris VII.

DAVID (Sophie) 1993a. *Les Unités polylexicales : éléments de description et reconnaissance automatique.* Thèse de doctorat en linguistique théorique et formelle. Paris : Université de Paris VII.

DAVID (Sophie) 1993b. « Remarques à propos du mode de construction des unités de forme NN ». *T.A.L.*, vol. 34, n° 2, p. 59-74.

DAVID (Sophie) et Pierre Plante 1990a. « De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes ». *ICO*, vol. 2, n° 3, p. 140-155.

DAVID (Sophie) et Pierre Plante 1990b. *Termino v.1.0™ : rapport de recherche.* Document de travail RDLC (Centre d'ATO.CI). Montréal : Université du Québec à Montréal [n.p.].

DAVID (Sophie) et Pierre Plante 1991. « Termino v. 1.0™ : principes et propriétés linguistiques » *in* Actes du colloque *Industries de la langue*, nov. 1990. Montréal : OLF et Société des traducteurs du Québec. P. 71-88.

DAVID (Sophie) et Maryse Souchard, 1995. « Analyse de discours et traitement automatique de données textuelles : le logiciel Termino ». *Cahiers de linguistique sociale.* (Actes du colloque « Recherches documentaires »), p. 61-76.

DAVID (Sophie), Najib Faraj, Robert Godin, Rokia Missaoui et Pierre Plante, 1996. « Analyse d'une méthode d'indexation automatique fondée sur une analyse syntaxique de texte ». *Revue canadienne des sciences de l'information et de bibliothéconomie*, vol. 21, n° 1, p. 1-21.

DELEDALLE (Gérard) 1990. *Lire Peirce aujourd'hui.* Bruxelles : De Boeck-Wesmael.

DE LIBERA (Alain) 1993. *La Philosophie médiévale.* Paris : Presses universitaires de France. (Collection Premier Cycle).

DE LIBERA (Alain) et Irène Rosier, 1992. « L'Analyse de la référence » *in Histoire des idées linguistiques.* Tome 2 : le développement de la grammaire occidentale. Liège : P. Mardaga. (Philosophie et langage). P. 137-158.

DESMET (Isabel) et Samy Boutayeb, 1993. « Terme et mot : proposition pour la terminologie ». *La Banque des mots*, n° 5 (numéro spécial), p. 6-32.

DEWESE (André) 1993. *Informatique documentaire.* Paris : Masson.

Dictionnaire de linguistique et des sciences du langage 1994. Paris : Larousse.

DUBOIS (Danièle) 1995. « Interrogation documentaire : recherche d'information ou gestion de connaissances ? » *Cahiers de linguistique sociale.* (Actes du colloque « Recherches documentaires »), p. 87-96.

DUBOIS (Danièle) et Lorenza Mondada, 1995. « Construction des objets de discours et catégorisation : une approche des processus de référenciation ». *TRANEL*, n° 23, p. 273-302.

DUCROT (Oswald) et Jean-Marie Schaeffer, 1995. *Nouveau dictionnaire encyclopédique des sciences du langage.* Paris : Seuil.

ECO (Umberto) 1985 [1979]. *Lector in fabula : le rôle du lecteur ou la coopération interprétative dans les textes narratifs.* Paris : Grasset & Fasquelle. (Le Livre de poche, Biblio-Essais ; 4098).

ECO (Umberto) 1992 [1990]. *Les Limites de l'interprétation.* Paris : Grasset & Fasquelle. (Le Livre de poche, Biblio-Essais ; 4192).

ECO (Umberto) 1994. *La Recherche de la langue parfaite dans la culture européenne.* Paris : Seuil. (Faire l'Europe).

ENDRES-NIGGEMEYER (Brigitte) 1989. « A procedural model of abstracting and some ideas for its implementation, » *in TKE'90, Terminology and Knowledge Engineering Applications.* Franckfurt : Indeks Verlag. P. 231-241.

ENGEL (Pascal) 1985. *Identité et référence : la théorie des noms propres chez Frege et Kripke.* Paris : Presses de l'École Normale Supérieure.

ESCARPIT (Roger) 1991. *L'Information et la communication : théorie générale.* Paris : Hachette (HU ; Communication).

EVERAERT-DESMEDT (Nicole) 1990. *Le Processus interprétatif : introduction à la sémiotique de Ch. S. Peirce.* Liège : P. Mardaga. (Philosophie et langage).

FAUCONNIER (Georges) 1984. *Espaces mentaux.* Paris : Éditions de Minuit.

FORMIGARI (Lia) 1992. « Le langage et la pensée », *in Histoire des idées linguistiques.* Tome 2 : le développement de la grammaire occidentale. Liège : P. Mardaga. (Philosophie et langage). P. 442-465.

FOUCAULT (Michel) 1969. *Archéologie du savoir.* Paris : Gallimard.

FOUCAULT (Michel) 1984. *Histoire de la sexualité (2) : l'usage des plaisirs.* Paris : Gallimard. (Bibliothèque des histoires).

FRADIN (Bernard) 1984. « Langue, discours, lexique ». *Linx,* n° 10, p. 159-165.

FRADIN (Bernard) et Jacqueline Léon, 1985. « Les abords de l'ambigu : approche d'une description automatique du sens lexical ». *Brises,* n° 7, p. 72-80.

FRADIN (Bernard) et Jean-Marie Marandin, 1979.« Autour de la définition : de la lexicographie à la sémantique ». *Langue française,* n° 43, p. 60-83.

FRANCKEL (Jean-Jacques) 1992. « De l'invariance opératoire à la polysémie : le sens du verbe *porter* ». *Cahiers de lexicologie,* vol. 61, n° 2, p. 18-39.

FREGE (Gottlob) 1971 [1892]. « Sens et dénotation ». *Écrits logiques et philosophiques.* Paris : Seuil. (L'Ordre philosophique). P. 102-126.

FUCHS (Catherine) 1982. *La Paraphrase.* Paris : Presses universitaires de France. (Linguistique nouvelle).

FUCHS (Catherine) sous la dir. de, 1993. *Linguistique et traitements automatiques des langues.* Paris : Hachette. (Supérieur).

FUGMANN (Robert) 1993. *Subject analysis and indexing : theoretical foundations and practical advice.* Franckfurt : Indeks Verlag.

GARDIN (Jean-Claude) 1967. « Recherches sur l'indexation automatique des documents scientifiques ». *Revue d'informatique et de recherche opérationnelle,* n° 5, p. 27-46.

GARDIN (Jean-Claude) 1974. « Analyse documentaire et théorie linguistique » *in Les Analyses du discours.* Neuchâtel : Delachaux et Niestlé. P. 120-168.

GARDIN (Jean-Claude) 1991. *Le Calcul et la raison : essai sur la formalisation du discours savant.* Paris : Éditions de l'École des Hautes Études en Sciences Sociales.

GARY-PRIEUR (Marie-Noëlle) 1994. *Grammaire du nom propre.* Paris : Presses universitaires de France. (Linguistique nouvelle).

GARY-PRIEUR (Marie-Noëlle) 1996. « Figurations de l'individu à travers différentes constructions du nom propre en français ». *Cahiers de praxématique*, n° 27, p. 57-72.

GRANGER (Gilles-Gaston) 1993. *La Science et les sciences.* Paris : Presses universitaires de France. (Que sais-je ? ; 2710).

GROLIER de (Éric) 1962. *Étude sur les catégories applicables aux classifications et codifications documentaires.* Paris : UNESCO.

GROLIER de (Éric) 1970. « Quelques travaux récents en matière de classification encyclopédique ». *Bulletin des bibliothèques de France*, n° 1, 25e année, p. 99-126.

GROLIER de (Éric) 1988. « Taxilogie et classification : un essai de mise au point et quelques notes de prospective ». *Bulletin des bibliothèques de France*, t. 33, n° 6, p. 468-489.

GUIMIER-SORBETS (Anne-Marie) 1993. « Des textes aux images : accès aux informations multimédias par le langage naturel ». *Documentaliste - Sciences de l'information*, vol. 30, n° 3, p. 127-134.

HABERMAS (Jürgen) 1973 [1968]. *La Technique et la science comme « idéologie ».* Paris : Gallimard. (Tel).

HEMPEL (Carl) 1996 [1972]. *Éléments d'épistémologie.* Paris : Masson & A. Colin. (Cursus ; Philosophie).

HODGE (Gail M.) 1992. *Automated support to indexing.* Philadelphia : National Federation of Abstracting and Information.

HUDRISIER (Henri) 1984. *L'Iconothèque, documentation audiovisuelle et banques d'images.* Paris : La Documentation française.

HUDRISIER (Henri) 1996. Communication faite à l'ENSSIB (Mars 1996 ; Villeurbanne).

JACOB (André) 1976. *Introduction à la philosophie du langage.* Paris : Gallimard. (Collection Idées ; 354. Philosophie).

JACOBI (Daniel) 1984. *Recherches sociolinguistiques et interdiscursives sur la diffusion et la vulgarisation des connaissances scientifiques.* Thèse pour le doctorat d'État. Besançon : Université de Besançon.

JACOBI (Daniel) 1987. *Textes et images de la vulgarisation scientifique.* Berne : Peter Lang.

JACOBI (Daniel) 1988. « La vulgarisation scientifique : thèmes de recherche » *in Vulgariser la science : le procès de l'ignorance* (sous la dir. de Daniel Jacobi et Bernard Schiele). Seyssel : Champ Vallon. P. 12-45.

JACOBI (Daniel) 1990. « Les séries superordonnées dans les discours de vulgarisation scientifique ». *Langages*, n° 98, p. 103-115.

JACOBI (Daniel) 1993. « Les terminologies et leur devenir dans les textes de vulgarisation scientifique ». *Didascalia*, n° 1, p. 69-83.

JACOBI (Daniel) 1994. « Le devenir des terminologies dans la presse de vulgarisation : le cas des sciences de la terre » [texte de la communication] *in Terminologie & Information*, rencontre du 22 juin 1994, n.p., [10 p.].

JANIK (Sophie) 1985. « Linguistique et sciences de l'information : réalités et perspectives d'une approche interdisciplinaire ». *Argus*, vol. 14, n° 3, p. 81-93.

JEANNERET (Yves) 1994. *Écrire la science*. Paris : Presses universitaires de France.

JONASSON (Kerstin) 1994. *Le nom propre : constructions et interprétations*. Louvain-la-Neuve : Éditions Duculot. (Champs linguistiques).

KAYSER (Daniel) 1987. « Une sémantique qui n'a pas de sens ». *Langages*, n° 87, p.33-45.

KENT (Berry), Luehrs et Perry 1955. « Operational criteria for designing information retrieval systems ». *American Documentation*, vol. 6, n° 2, p. 93-101.

KERBRAT-ORRECCHIONI (Catherine) 1980. *L'énonciation : de la subjectivité dans le langage*. Paris : A. Colin. (Linguistique).

KERBRAT-ORRECCHIONI (Catherine) 1996. « Texte et contexte ». *SCOLIA*, n° 6, p. 39-60.

KERLEROUX (Françoise) 1996. *La Coupure invisible : études de syntaxe et de morphologie*. Lille : Presses universitaires du Septentrion.

KLEIBER (Georges) 1981. *Problèmes de référence : descriptions définies et noms propres*. Paris : Klincksieck. (Recherches linguistiques).

KLEIBER (Georges) 1994. *Nominales : essai de sémantique référentielle*. Paris : Armand Colin. (Linguistique).

KOCOUREK (Rostislav) 1991a. *La Langue française de la technique et de la science : vers une linguistique de la langue savante*. Wiesbaden : Brandsletter.

KOCOUREK (Rostislav) 1991b. « Textes et Termes ». *Méta*, vol. XXXVI, n° 1, p. 71-76.

KOLMAYER (Élisabeth) 1995. « Représentations de la situation d'interrogation : quelques approches en sciences de l'information ». *Cahiers de linguistique sociale*. (Actes du colloque « Recherches documentaires »), p. 27-33.

KOLMAYER (Élisabeth) 1997. *Contribution à l'analyse des processus cognitifs mis en jeu dans l'interrogation d'une banque de données documentaires*. Doctorat de psychologie. Paris : Université Paris V-René Descartes.

KRIPKE (Saul) 1982 [1972]. *La Logique des noms propres*. Paris : Éditions de Minuit.

KURAMOTO (Hélio) 1995. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux*. DEA en sciences de l'information. Villeurbanne : École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

KURAMOTO (Hélio) [thèse en cours]. Villeurbanne : École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

LAINÉ-CRUZEL (Sylvie) 1994. « Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs ». *Documentaliste-Sciences de l'information*, vol. 31, n° 3, p. 143-147.

LALLICH-BOIDIN (Geneviève) 1986. *Analyse syntaxique automatique du français écrit : applications à l'indexation automatique*. Thèse de doctorat. Grenoble : Université de Grenoble II .

LANCASTER (F.W.) 1991. *Indexing in theory and practise*. Urbana : Univ. of Illinois Press.

LANE (Philippe) 1992. *La Périphérie du texte*. Paris : Nathan. (Fac).

LARDY (Jean-Pierre) 1994. *Les Outils de recherche dans le World Wide Web* [support de cours ; n.p.].

LAROUK (Omar) 1994. *Extraction des connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation)*. Thèse de doctorat. Lyon : Université Lyon I-Claude Bernard.

LATOUR (Bruno) 1996. « Ces réseaux que la raison ignore : laboratoires, bibliothèques, collections » *in Le Pouvoir des bibliothèques : la mémoire du livre en Occident* (sous la dir. de Marc Baratin et de Christian Jacob). Paris : Albin Michel. (Histoire). P. 23-46.

LAZSLO (Pierre) 1993. *La Vulgarisation scientifique*. Paris : Presses universitaires de France. (Que sais-je ? ; 2722).

LE COADIC (YVES F.) 1994. *La Science de l'information*. Paris : Presses universitaires de France. (Que sais-je ? ; 2873)

LE COADIC (YVES F.) 1997. « Science de l'information » *in Dictionnaire encyclopédique de l'information et de la documentation*. Paris : Nathan. (Réf.). P. 516-523.

LE CROSNIER (Hervé) 1996. « Les bibliothécaires et le réseau » *in Les Nouvelles technologies dans les bibliothèques*, sous la dir. de Michèle Rouhet. Paris : Éditions du Cercle de la librairie. (Collection Bibliothèques). P. 349-372.

LE GUERN (Michel) 1984. « Les Descripteurs d'un système documentaire : essai de définition ». *Condenser*, Suppl. 1, p. 163-169.

LE GUERN (Michel) 1989. « Sur les relations entre terminologie et lexique ». *Méta*, vol. XXXIV, n° 3, p. 340-343.

LE GUERN (Michel) 1991a. « Un Analyseur morpho-syntaxique pour l'indexation automatique ». *Le Français moderne*, 1 (59), p. 22-35.

LE GUERN (Michel) 1991b. « Pour une approche logique de l'attribut grammatical » *in À la recherche de l'attribut*. Lyon : Presses universitaires de Lyon. P. 71-81.

LE GUERN (Michel) 1994. « Traitement automatique et variation linguistique : la syntaxe des titres », *in Opérateurs et constructions syntaxiques*. Paris : Presses de l'École Normale Supérieure. P. 75-81.

LE GUERN (Michel) 1997. « Linguistique » *in Dictionnaire encyclopédique de l'information et de la documentation*. Paris : Nathan. (Réf.). P. 375-379.

LE GUERN (Odile) 1989. « Images et bases de données ». *Bulletin des bibliothèques de France*. Tome 34, n° 5, p. 422-435.

LE LOARER (Pierre) 1994. « Indexation automatique, recherche d'information et évaluation » *in Le Traitement électronique du document*. Cours INRIA, Aix-en-Provence, 3-7 octobre 1994. Paris : ADBS. P. 149-201.

LE MOAL (Jean-Claude) 1997. « Logiciel documentaire » *in Dictionnaire encyclopédique de l'information et de la documentation*. Paris : Nathan. (Réf.). P. 380-384.

LE MOIGNE (Jean-Louis) 1995. *Les Épistémologies constructivistes*. Paris : Presses universitaires de France. (Que sais-je ? ; 2969).

LERAT (Pierre) 1995. *Les langues spécialisées*. Paris : Presses universitaires de France. (Linguistique nouvelle).

LESPINASSE (Karine) 1997. « TREC : une conférence pour l'évaluation des systèmes d'information ». *Documentaliste-Sciences de l'information*, vol. 34, n° 2, p. 74-81.

MAI CHAN (Lois), Philip A. Richmond and Elaine Svenomius (eds.), 1985. *Theory of subject analysis : a sourcebook*. Littleton (Colorado) : Libraries Unlimited Inc.

MALRIEU (Denise) 1992. « Les Apports d'une étude différentielle de la demande bibliographique pour la modélisation des utilisateurs ». *Intellectica*, n° 15, p. 187-214.

MALRIEU (Denise) 1994. « Genre textuel, surlignages et marques linguistiques d'importance ». *Cahiers de lexicologie*, p. 123-140.

MANIEZ (Jacques) 1976-1977. *Le rôle de la syntaxe dans les systèmes de recherche documentaire*. Tome 1 : aspects linguistiques. Tome 2 : étude critique de quelques SRD. Thèse de doctorat. Dijon : IUT de Dijon.

MANIEZ (Jacques) 1987. *Les Langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires*. Paris : Éditions d'Organisation. (Système d'information et de documentation).

MANIEZ (Jacques) 1993. « L'Évolution des langages documentaires ». *Documentaliste-Sciences de l'information*, vol. 30, n° 4-5, p. 254-259.

MARANDIN (Jean-Marie) 1979. « Problèmes d'analyse du discours : essai de description du discours français sur la Chine ». *Langages*, n° 55, p. 17-88.

MARANDIN (Jean-Marie) 1984. « Mais qu'est-ce que Socrate a au juste à voir avec la sagesse ? ». *Linx*, n° 10, p. 51-55.

MARANDIN (Jean-Marie) 1988. « À propos de la notion de thème de discours. Éléments d'analyse dans le récit ». *Langue française*, n° 78, p. 67-87.

MARANDIN (Jean-Marie) 1990. « Le lexique mis à nu par ses célibataires : stéréotype et théorie du lexique » *in La Définition* (actes du colloque CELEX). Paris : Larousse. P. 284 -291.

MARANDIN (Jean-Marie) 1992a [présentation]. « L'Individualité lexicale ». [Présentation des textes des communications]. *Cahiers de lexicologie*, vol. 61, n° 2, p. 6-10.

MARANDIN (Jean-Marie) 1992a. « Il y a de la synonymie ». *Cahiers de lexicologie*, vol. 61, n° 2, p. 39-57.

MARANDIN (Jean-Marie) 1992b. « La perception syntaxique ». *Le Gré des langues*, n° 4, p. 64-91.

MARANDIN (Jean-Marie) 1993. « Syntaxe, discours, du point de vue de l'analyse de discours ». *Histoire, Épistémologie, Langage*, tome 15, fasc. II, p. 155-177.

MARANDIN (Jean-Marie) 1997. *Perception syntaxique et constructions syntaxiques*. Mémoire d'habilitation. Paris : Université Paris VII-Denis Diderot.

MARIETTI-K. (Angèle) 1985 [1974]. *Michel Foucault : archéologie et généalogie*. Paris : Grasset & Fasquelle. (Le Livre de poche, Biblio-Essais ; 4036).

MENON (Bruno) 1988. « Indexation automatique et intelligence artificielle : quelques questions de stratégie », *in Image et intelligence artificielle dans l'information scientifique et technique*. Cours INRIA des 6-10 juin 1988 dirigé par Christian Bornes. Le Chesnay : INRIA. P. 145-171.

METZGER (Jean-Paul) 1988. *Syntagmes nominaux et information textuelle : reconnaissance automatique et représentation*. Thèse de doctorat. Lyon : Université de Lyon I-Claude Bernard.

METZGER (Jean-Paul) 1997. « Logique » *in Dictionnaire encyclopédique de l'information et de la documentation*. Paris : Nathan. (Réf.). P. 385-390.

MICHEL (Jean) 1997. « Internet » *in Dictionnaire encyclopédique de l'information et de la documentation.* Paris : Nathan. (Réf.). P. 361-363.

MIÈGE (Bernard) 1993-1994. « Les Étapes de la pensée communicationnelle » [chronique en trois volets]. *Sciences de la société,* n° 29, mai 1993, p. 198-210 ; n° 30, octobre 1993, p. 193-204 ; n° 31, février 1994, p. 187-196.

MILNER (Jean-Claude) 1976. « Réflexions sur la référence ». *Langue française,* n° 30, p. 63 -73.

MILNER (Jean-Claude) 1978. *De la syntaxe à l'interprétation : quantités, insultes, exclamations.* Paris : Seuil. (Travaux linguistiques)

MILNER (Jean-Claude) 1989. *Introduction à une science du langage.* Paris : Seuil, 1989. (Les Travaux)

MILNER (Jean-Claude) 1992. « Individualité lexicale et sémantique ». *Cahiers de lexicologie,* vol. 61, n° 2, p. 10-18.

MORTUREUX (Marie-Françoise) 1983. *La Vulgarisation scientifique au XVIII^e siècle à travers l'œuvre de Fontenelle.* Paris : Didier-Érudition.

MORTUREUX (Marie-Françoise) 1984. « La dénomination : approche socio-linguistique ». *Langages,* n° 76, p. 95-112.

MORTUREUX (Marie-Françoise) 1988. « La vulgarisation scientifique : parole médiane ou dédoublée » *in Vulgariser la science : le procès de l'ignorance* (sous la dir. de Daniel Jacobi et Bernard Schiele). Seyssel : Champ Vallon. P. 118-147.

MORTUREUX (Marie-Françoise) 1993. « Comment peut-on définir la propriété d'un mot ? » *in Parcours linguistiques de discours spécialisés.* Actes de colloque des 23-24-25 septembre 1992 à l'université Paris IV-Sorbonne. Berne : Peter Lang. (Sciences pour la communication). P. 3-9.

MORTUREUX (Marie-Françoise) 1995. « Les Vocabulaires scientifiques et techniques » *in Les Enjeux des discours spécialisés.* Paris : Presses de la Sorbonne Nouvelle. (Les Carnets du CEDISCOR ; 3). P. 13-25.

MORTUREUX (Marie-Françoise) et Gérard Petit 1989. « Fonctionnement du vocabulaire dans la vulgarisation et problèmes de lexique ». *DRLAV,* n° 40, p. 41-62.

MOUNIN (Georges) 1963. *Les Problèmes théoriques de la traduction.* Paris : Gallimard. (Tel).

MOUNIN (Georges) sous la dir. de, 1974. *Dictionnaire de la linguistique.* Paris : Presses universitaires de France.

MUSTAFA-ELHADI (Widad) 1989. *La Terminologie arabe des télécommunications : faits de variation.* Thèse de doctorat en sciences du langage. Lyon : Université Lyon II-Lumière.

MUSTAFA-ELHADI (Widad) 1992. « La Contribution de la terminologie à la conception théorique des langages documentaires et à l'indexation des documents ». *Méta,* vol. XXXIV, n° 3, p. 465-473.

NEET (Hanna E.) 1989. *À la recherche du mot-clé : analyse documentaire et indexation alphabétique.* Genève : IES. (Les Cours de l'IES).

NEF (Frédéric) 1991. *Logique, langage et réalité.* Paris : Éditions universitaires.

PEIRCE (Charles S.) 1978. *Écrits sur le signe* [textes rassemblés, traduits et commentés par Gérard Deledalle]. Paris : Seuil. (L'Ordre philosophique).

PERRON (Jean) 1988. « Le dépouillement terminologique assisté par ordinateur ». *Terminogramme*, n° 46, p. 24-31.

PERRON (Jean) 1991. « Présentation du progiciel de dépouillement terminologique assisté par ordinateur : Termino » *in* Actes du colloque *Industries de la langue* (nov. 1990). Montréal : OLF et Société des traducteurs du Québec. P. 715-755.

POMART (Paul-Dominique) et Éric Sutter 1997. « Indexation » *in Dictionnaire encyclopédique de l'information et de la documentation*. Paris : Nathan. (Réf.). P. 284-287.

PORTELANCE (Christine) 1989. « Syntagmes et paradigme ». *Méta*, vol. XXXIV, n° 3, p. 398-404.

PRINCE (Violaine) 1996. *Vers une informatique cognitive dans les organisations : le rôle central du langage*. Paris : Masson. (Sciences cognitives).

PUTNAM (Hilary) 1990a. *Représentation et réalité*. Paris : Gallimard.

PUTNAM (Hilary) 1990b (trad.) [1970]. « La sémantique est-elle possible ? » *in La Définition* (actes du colloque CELEX). Paris : Larousse. P. 292-304.

QUINN (Brian) 1994. « Recent Theoretical Approaches in Classification and Indexing ». *Knowledge Organisation*, vol. 21, n° 3, p. 140-147.

RAICHVARG (Daniel) et Jean Jacques 1991. *Savants et ignorants : une histoire de la vulgarisation des sciences*. Paris : Seuil.

RANGANATHAN (S.R.) assisted by M.A. Gopinath, 1967. *Prolegomena to Library Classification*. Bombay : Asia Publishing House.

RANGANATHAN (S.R.) 1976. *Colon Classification*. Bombay : Asia Publishing House.

RASTIER (François), Marc Cavazza et Anne Abeillé, 1994. *Sémantique pour l'analyse : de la linguistique à l'informatique*. Paris : Masson. (Sciences cognitives).

RÉCANATI (François) 1983. « La Sémantique des noms propres : remarques sur la notion de " désignateur rigide " ». *Langue française*, n° 57, p. 106-118.

REY (Alain) 1992. *La Terminologie. Noms et Notions*. Paris : Presses universitaires de France. (Que sais-je ? ; 1780).

REY-DEBOVE (Josette) 1978. *Le Métalangage : étude linguistique du discours sur le langage*. Paris : Le Robert. (L'Ordre des mots).

RIAO (Recherche d'information assistée par ordinateur) 1994. « Recherche et développement des systèmes d'information multimédia : synthèse de la conférence ». Quatrième conférence, 11-13 octobre 1994. *Informatique documentaire*, n° 57-58.

RICŒUR (Paul) 1971. « Événement et sens dans le discours » [texte inédit] *in Ricœur* (Michel Philibert). Paris : Seghers. (Philosophes de tous les temps). P. 177-187.

RICŒUR (Paul) 1975. *La Métaphore vive*. Paris : Seuil.

RICŒUR (Paul) 1986. *Du texte à l'action. Essais d'herméneutique*. Tome II. Paris : Seuil.

RÔLE (François) 1993. « De la lettre au sens : les recherches en texte intégral ». *Documentaliste - Sciences de l'information*, vol. 30, n° 3, p. 136-146.

ROUAULT (Jacques) 1987. *Linguistique automatique : applications documentaires*. Berne : Peter Lang.

SALAÜN (Jean-Michel) 1991. [L'Information documentaire] *in Communication et nouvelles technologies*, actes de colloque décembre 1991, sous la coord. de Claire Bélisle. *Chemins de la recherche*, n° 16, p. 139-142.

SALTON (Gerard) 1986. « On the use of term association in automatic information retrieval » *in Coling '86*, 11th international conference on computational linguistics, 25-29 août 1986, Bonn. P. 380-386.

SALTON (Gerard) 1988. « Automatic text indexing using complex identifiers » *in Proceedings of the ACM*, conference on document processing system, Santa Fe (New Mexico), december 5-9 1988. P. 135-144.

SALTON (Gerard) 1990. « On the application of syntactic methodologies in automatic analysis ». *Information Processing and Management*, vol. 26, n° 1, p. 73-92.

SALVAN (Paule) 1972. *Esquisse de l'évolution des systèmes de classification.* Paris : École Nationale Supérieure des Bibliothèques.

SARACEVIC (Tefko) 1992. « Information science : origin, evolution and relations », *in Conceptions of library and information science : historical, empirical and theoretical perspectives.* London : Taylor Graham. P. 5-25.

SAUSSURE de (Ferdinand) 1973. *Cours de linguistique générale*, édition critique de Tullio de Mauro. Paris : Payot.

SELOSSE (Philippe) 1996. « À propos de " nom propre et nomination " ». *Le Français moderne*, LXIV, n° 2, p. 107-224.

SFBA (Société Française de Bibliométrie Appliquée) 1995. « Les Systèmes d'information élaborée » Actes des journées 30 mai-2 juin 1995, Île Rousse. *Revue française de bibliométrie*, n° 14.

SIDHOM (Saadi) [thèse en cours]. *Conception d'une bibliothèque virtuelle via Internet.* Villeurbanne : École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

SIMONOT (Françoise) 1993. *Utilisation d'un logiciel d'extraction terminologique à des fins d'indexation.* Mémoire de maîtrise en science de l'information et de la documentation. Paris : Université de Paris I.

SOUCHARD (Maryse) 1989. *Le Discours de presse : l'image des syndicats au Québec (1982-1983).* Québec (Montréal) : Éditions du Préambule. (L'Univers des discours).

SPARCK JONES (Karen) ed., 1981. *Information Retrieval Experiment.* London ; Boston ; Sydney : Butterworths.

SUTTER (Éric) 1997a. « La recherche documentaire » *in Dictionnaire encyclopédique de l'information et de la documentation.* Paris : Nathan. (Réf.). P. 486-488.

SUTTER (Éric) 1997b. « Document primaire » *in Dictionnaire encyclopédique de l'information et de la documentation.* Paris : Nathan. (Réf.). P. 194.

TÊTU (Jean-François) 1997. « Science de la communication » *in Dictionnaire encyclopédique de l'information et de la documentation.* Paris : Nathan. (Réf.). P. 513-516.

THIBAUD (P.) 1989. « Nom propre et individualisation chez Peirce ». *Dialectica*, vol. 43, n° 4, p. 373-386.

TURNER (William A.) 1990. « Perspectives de l'indexation assistée par ordinateur » *in Indexation automatique en France : état de la recherche, problèmes rencontrés et analyse de produits disponibles.* Paris : Bureau Van Djik. [8 p.].

TURNER (William A.) 1994. « Penser l'entrelacement de l'Humain et du Technique : les réseaux hybrides d'intelligence » *in Pour une nouvelle économie du savoir.* Rennes : Presses universitaires de Rennes. (Dossier du GIRSIC n°1). P. 21-49.

TYVAERT (Jean-Emmanuel) 1994. « Initialisation de la référence actuelle et organisation différentielle de la référence virtuelle ». *SCOLIA*, n° 1, p. 41-53.

VAN HOOLAND (Michelle) 1995. « Indexation : des données du monde aux données de sens ». *Cahiers de linguistique sociale.* (Actes du colloque « Recherches documentaires »), p. 109-120.

VAN SLYPE (Georges) 1987. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Paris : Éditions d'Organisation. (Système d'information et de documentation).

VARET (Gilbert) et Marie-Madeleine Varet, 1995. *Maîtriser l'information à travers sa terminologie.* Paris : Les Belles Lettres.

VÉRON (Éliseo) 1990. *Espaces du livre : perception et usages de la classification et du classement en bibliothèque.* Paris : Bibliothèque Publique d'Information. (Études et recherche).

VICKERY (B.C.) 1963. *La Classification à facettes : guide pour la construction et l'utilisation de schémas spéciaux.* Paris : Gauthier-Villars.

VIDALENC-SABOURIN (Isabelle) 1989. *Traitement automatique des anaphores en français : étude linguistique préalable.* Thèse de doctorat en sciences de l'information et de la communication. Lyon : Université Lyon II-Lumière.

Vocabulaire de la documentation 1987. Paris : AFNOR. (Les Dossiers de la normalisation)

WILLEMS (Emilio) 1970. *Dictionnaire de la sociologie.* Paris : Librairie Marcel Rivière & Cie.

WINOGRAD (Terry) et Fernando Flores, 1989. *L'intelligence artificielle en question.* Paris : Presses universitaires de France.

ZIMMERMANN (Francis) 1989. « Les Théories de la signification » *in Histoire des idées linguistiques.* Tome 1 : la naissance des métalangages en Orient et en Occident. Liège : P. Mardaga. (Philosophie et langage). P. 401-416.

TABLE DES MATIÈRES

ADBS EDITIONS

Collection Sciences de l'information, série Recherches et documents

dirigée par Serge Cacaly

Extrait du catalogue

Catalogue complet : http://www.adbs.fr

- *Petit guide d'accès à l'information juridique française : pratique de la recherche documentaire juridique*, par Stéphane Cottin et Sophie Moyret. 2000
- *Communiquer les publications multimédia en bibliothèque et centre de documentation : description des systèmes de gestion des ressources électroniques*, étude rédigée par Marc Maisonneuve et Annie Gourdier. 2000
- *L'évolution de la fonction Information-Documentation : résultats de l'enquête ADBS 1999*, rapport rédigé par Benoît Roederer. 2000
- *Terminologie et documentation : pour une meilleure circulation des savoirs*, par Maryvonne Holzem. 1999
- *Comment concevoir un service web : de la théorie à la pratique*, par Arnaud Le Guelvouit. 1999
- *Les enjeux du management de l'information dans les organisations : usages, outils, techniques*, Observatoire des nouvelles technologies de l'information, par les étudiants du DESS Stratégies de l'information et de la documentation de l'Université de Lille 3 ; préface de Dominique Cotte. 1999
- *Recherche d'information sur l'Internet : outils et méthodes*, par Jean-Pierre Lardy. Sixième édition mise à jour en mai 1999
- *S'informer en Bourgogne : répertoire 1999 des centres de documentation*, par le groupe régional Bourgogne de l'ADBS. 1999
- *Répertoire des centres de documentation Auvergne-Limousin*, par le groupe régional Auvergne de l'ADBS. 1999
- *Le droit de copie en questions*, par la commission Droit de l'information de l'ADBS. 1998
- *Diffuser la documentation via Intranet et Internet : description des serveurs Web associés aux systèmes de gestion documentaire et de bibliothèque*, étude rédigée par Michèle Lénart, Nadia Bony et Marc Maisonneuve. 1998
- *Diffuser sur Internet le catalogue de la bibliothèque : description des serveurs Web associés aux systèmes de gestion documentaire et de bibliothèque*, étude rédigée par Michèle Lénart, Nadia Bony et Marc Maisonneuve. 1998
- *S'informer en Bretagne : guide 1998 des sources documentaires*, par le groupe régional Bretagne de l'ADBS. 1998
- *L'état des nouvelles technologies de l'information en 1998* : Observatoire des nouvelles technologies de l'information, par les étudiants du DESS Systèmes informationnels et documentaires de l'Université de Lille 3 ; préface de Dominique Cotte. 1998
- *Stratégies informationnelles et valorisation de la recherche scientifique publique*, ouvrage coordonné par Françoise Renzetti ; préface de Guy Aubert. 1998
- *Modèles de communication et stratégies d'entreprises : problème d'organisation ou problème de management ?* Actes du colloque international TransInfo 96 - CNISF, FMOI, CNAM ; ouvrage coordonné par Liliane Vézier et Danièle Bretelle-Desmazières. 1997

Collection Sciences de l'information

Série Recherches et documents

Collection dirigée par Serge Cacaly

Cette collection est consacrée aux principaux aspects des Sciences de l'information : méthodes et techniques, usages et usagers, secteurs et applications. La série Recherches et documents vise à mettre rapidement entre les mains des professionnels de l'information et de la documentation des textes de «littérature grise» en prise sur leurs pratiques et leurs reflexions : on y trouvera des travaux de recherche (thèses et mémoires de fin d'études), des rapports administratifs, des outils de travail, des documents issus de colloques et de journées d'étude...

L'indexation documentaire est le plus souvent appréhendée par les professionnels de l'information dans sa seule dimension instrumentale : c'est en se focalisant sur sa finalité - la recherche d'information - qu'elle se définit d'abord. Si le choix d'un tel angle d'approche se comprend aisément dans le cadre des pratiques professionnelles, il se révèle réducteur dès lors que l'indexation devient l'objet d'autres problématiques, notamment celles de l'évolution des technologies de l'information.

La recherche d'où est issu cet ouvrage s'est intéressée à dégager les spécificités, les caractéristiques, les propriétés de l'indexation et, pour ce faire, à en étudier les fondements théoriques par le biais d'interrogations issues de la linguistique. La première partie est consacrée aux « Problèmes théoriques de l'indexation » : la question du lexique et celle de la référence y sont examinées du point de vue des professionnels et de celui des linguistes. La seconde partie est une « Contribution aux fondements théoriques de l'indexation » : elle s'attache à en faire émerger les aspects « discursifs », tant au niveau du processus que du résultat. Cette approche permet de situer l'indexation dans le cadre plus large des pratiques de diffusion des connaissances, la rapprochant ce faisant des problématiques de la vulgarisation scientifique.

Ce nouveau positionnement montre qu'il est possible d'envisager l'indexation dans le rapport particulier qu'elle entretient avec les textes eux-mêmes et les unités linguistiques qui les constituent : une tentative de « réconciliation » entre la langue et la pratique professionnelle de l'indexation qui devrait ouvrir des perspectives aux praticiens comme aux chercheurs.

ADBS
EDITIONS

25, rue Claude Tillier – F - 75012 Paris - Tél. : 01 43 72 25 25 – Télécopie 01 43 72 30 41
Mél. adbs@adbs.fr – Internet http://www.adbs.fr

ISBN 2-84365-042-9 – ISSN 1159-7666 – Prix : 180 F (27,44 euros)