«Indexation discursive» *versus* «indexation lexicale» Éléments de définition

Sur la base de l'opposition linguistique entre lexique et discours, deux types d'indexation – lexicale et discursive – sont distingués.

L'indexation lexicale est évaluée au regard de la notion de contexte documentaire: elle se révèle adéquate uniquement dans des contextes documentaires reposant sur la présence d'intermédiaires (humain et textuel).

L'indexation discursive est ensuite définie dans ses principales composantes: l'indexation devient un choix de textes (et non plus de mots) et le descripteur une unité synaptique identifiable uniquement par une procédure automatique.

Termes-clés:
indexation; lexique; discours;
corpus; synapsie.

Introduction

Cette contribution ⁽¹⁾ se propose de penser les enjeux actuels de l'indexation en exploitant l'opposition établie en linguistique entre lexique et discours (Le Guern 1989).

Du point de vue de cette opposition, on distingue deux types d'indexation: indexation lexicale et indexation discursive. On montre que ces deux types d'indexation ne peuvent être pratiqués indifféremment: chacun relève d'un contexte documentaire déterminé, qu'il importe de caractériser.

C'est ainsi que, si l'indexation lexicale a été jusqu'à présent adéquate à la plupart des contextes documentaires que nous connaissons, il n'est pas sûr qu'elle révèle la même pertinence dans les contextes documentaires actuels, sur l'Internet par exemple ⁽²⁾.

L'évaluation de l'indexation lexicale que nous proposons, à travers la notion de contexte documentaire, nous amène à discuter des pratiques d'indexation que l'on nommera «normatives», en faisant référence à la norme Z 47-102 (Afnor 1978): ces pratiques, liant indissociablement indexation et langages documentaires, relèvent de l'indexation lexicale. Elles tendent à tenir un rôle de «modèle», au double sens de modèle théorique et de modèle de référence, alors qu'elles ne rendent pas compte de tous les types d'indexation: de quel statut épistémologique relèvent alors les

normes professionnelles dont elles se revendiquent?

Notre approche distinguant deux types d'indexation permet de dissocier le processus d'indexation de son instrument privilégié, le langage documentaire:

– La prépondérance du lexique en indexation, marquée par un recours massif aux langages documentaires, devient caractéristique de pratiques d'indexation spécifiques, contraintes par un environnement technique historiquement situé: l'utilisation d'un lexique reste-t-elle encore définitoire du processus d'indexation?

– Il devient possible de penser des pratiques d'indexation situées non plus en aval – du côté du choix des mots – mais en amont – du côté du choix des textes et de leur organisation discursive. Cette indexation, que l'on propose de nommer «discursive», permet-elle d'approcher plus adéquatement le processus d'indexation lui-même?

1 Indexation lexicale et indexation discursive: fondement de l'opposition

L'opposition lexique/discours établie en linguistique permet de poser

(1) Qui doit beaucoup à Sophie David, que je tiens à remercier ici.

(2) Nous adoptons ici l'hypothèse, sans la discuter, qu'Internet constitue un «espace documentaire»: la question est cependant plus complexe, (Jacquesson et Rivier 1999: chap. 15).

une double problématique qui intéresse directement les faits d'indexation : la problématique de la référence d'une part (1.1) et celle de la thématization d'autre part (1.2).

1.1 Problématique de la référence

La référence peut être entendue comme la « propriété d'un signe linguistique lui permettant de renvoyer à un objet extra-linguistique, réel ou imaginaire » (*Dictionnaire de linguistique et des sciences du langage*, 1994).

Michel Le Guern (1991 : 22) a montré que la problématique de la référence intéresse directement les faits d'indexation : « 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 ». Si on utilise des mots en indexation, ce n'est pas pour en donner le sens mais pour permettre de désigner des objets.

Or, d'un point de vue linguistique, toutes les unités ne sont pas référentielles, on distingue :

– Les unités du lexique – les unités lexicales (Milner 1989) – sont des unités linguistiques pourvues d'un

sens mais n'ayant pas de référence : ce sont des unités qui ne peuvent permettre, à elles seules, de désigner un objet du monde ;

– Les unités de discours sont, elles, des unités référentielles, susceptibles de désigner des objets du monde. Pour la catégorie nominale, les unités de discours sont les groupes nominaux : ils se construisent généralement ⁽³⁾ sur la base d'une unité lexicale.

Par exemple, *école* est une unité lexicale qui a un sens mais pas de référence ; c'est *école* insérée dans un groupe nominal, *l'école*, qui peut désigner un objet extra-linguistique.

Du point de vue de cette opposition, on dira que l'indexation normative, qui repose sur l'utilisation des langages documentaires, est lexicale : les descripteurs sont tous des unités lexicales de type nominal. Rappelons sur ce point que la norme Z 47-100 (Afnor 1981) stipule expressément que les descripteurs doivent être des « substantifs ».

Par opposition, on dira que les pratiques d'indexation qui exploitent les groupes nominaux issus des documents relèvent de l'indexation discursive.

On constate donc que l'indexation normative repose, d'un point de vue linguistique, sur un paradoxe : elle utilise des unités lexicales, sans référence, pour construire des référents.

En fait, le fonctionnement référentiel des unités lexicales est assuré, dans l'indexation normative, par les relations qui régissent l'organisation des termes dans les langages documentaires ⁽⁴⁾. Ces relations permettent de stabiliser la référence des unités lexicales utilisées : placer par exemple (*Thésaurus de l'Unesco* 1995) le terme « laser » sous la dépendance hiérarchique du terme « instrument de communication » exclut d'emblée la construction de l'objet « laser » utilisé en ingénierie médicale.

Contrairement à la construction référentielle « naturelle », qui va du sens à la référence, on peut dire que l'indexation normative propose une construction référentielle « artificielle » où la référence est première, où l'« accord de désignation » ⁽⁵⁾ supprime l'accord de signification.

Les deux types d'indexation – lexicale et discursive – ne s'opposent pas uniquement sur la façon dont la référence est construite : ils s'opposent également sur la façon dont le processus de thématization se réalise.

1.2 Problématique de la thématization

La thématization (l'identification des thèmes d'un document) est centrale en indexation : l'indexation consiste en effet, comme le note la norme (Afnor 1978 : 225), à « indiquer brièvement, sous forme concise, la teneur du document ».

Les études contemporaines en linguistique et plus particulièrement en analyse de discours (Marandin 1988 et 1997) ont permis de renouveler les approches de la thématization en distinguant la construction du thème, qui est de nature discursive, et sa formulation, qui est de nature lexicale : on retrouve ici l'opposition lexicale/discours déjà en jeu dans la référenciation.

Dans l'approche de Marandin, le thème d'un discours – le fait qu'un mot représente le contenu d'un texte – ne signifie plus que c'est le sens d'un mot qui indique la thématization d'un texte ; c'est plutôt l'interprétation – discursive – de ce mot qui construit la cohérence thématique d'un texte : ce que Barthes (1970 : 14) appelle « faire séquence » :

« 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

(3) Nous n'abordons pas ici le cas des noms propres.

(4) On distingue généralement (Afnor 1981) cinq types de relations dans les langages documentaires : relation hiérarchique, relation d'équivalence, relation associative, relation définitoire, relation catégorielle.

(5) On entend par « accord de désignation » le fait que, dans un langage documentaire, les professionnels se mettent d'accord pour utiliser tel terme pour désigner tel objet : cet accord est scellé par l'inscription de ce terme dans une relation.

développe au rythme de la nomination qui se cherche et se confirme».

Le thème n'est donc plus, dans cette approche, un nom sans lien au texte, issu du seul lexique, attribué de l'extérieur une fois pour toutes. Il est plutôt un «terme textuel», un nom tissé à même la trame du texte: il est fruit d'une interprétation: «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», (Marandin 1997: 21).

Reste que le thème, discursivement construit, se donne toujours et uniquement sous une forme lexicale: il se crée ainsi une illusion d'optique ou encore un «effet d'eschérisme» qui peut donner l'impression que c'est bien un nom qui représente thématiquement un texte alors que c'est le nom lui-même qui participe à la construction thématique du texte (Marandin 1988).

Cette approche de la thématization, qui dédouble la problématique en deux phases, construction discursive et formulation lexicale du thème, permet, à nouveau, de distinguer deux types d'indexation:

- L'indexation lexicale est celle qui se situe au niveau final de la formulation lexicale du thème. Dans ce cas, l'indexation interrompt le processus de thématization et fixe le texte dans une seule thématique ⁽⁶⁾; le choix du thème et le nom qui lui est donné sont ici le fait de l'indexeur;
- L'indexation discursive est celle qui se situe au niveau initial de la construction du thème. Dans ce cas, le thème est, au moment de l'indexation, inachevé du point de vue interprétatif; l'indexation consiste à maintenir les différents thèmes possibles, laissant ouverts, à l'utilisateur, tous les parcours interprétatifs: c'est donc ici lui et non plus l'indexeur qui thématise, c'est-à-dire qui achève une lecture et la nomme.

On reconnaîtra, dans le premier type d'indexation – lexicale –, les

pratiques normatives de l'indexation: les langages documentaires tiennent précisément un rôle de filtre thématique sur un corpus documentaire: ce sont eux qui permettent de stabiliser le texte dans un thème. Les langages documentaires permettent en effet, quand ils sont spécialisés (type thésaurus), de ne «voir» d'un document qu'une thématique spécifique et, quand ils sont généraux (type classifications), de «voir» toujours la même thématique dans des documents différents.

Thématiser revient, dans le cas de l'indexation lexicale, à retrouver dans un texte des unités déjà interprétées, c'est-à-dire des thèmes prédéfinis dans un langage documentaire.

Par opposition, thématiser consistera, dans le cas de l'indexation discursive, à permettre la construction des unités d'interprétation dont le texte dispose, sans restreindre les thèmes possibles.

1.3 Indexation lexicale ou indexation discursive: un simple choix?

L'opposition lexicale/discours empruntée à la linguistique permet:

- D'une part d'identifier, dans les pratiques d'indexation les plus courantes (pratiques que nous avons dénommées normatives), une prédominance du lexique à la fois dans le type d'unités retenues comme descripteurs (unités lexicales sans référence) et dans le point de vue porté sur la thématization (point de vue final de formulation lexicale du thème);
- D'autre part de disposer d'une alternative: une indexation, que l'on propose de nommer «discursive», peut être pensée, qui se singularise par le type d'unités considérées comme descripteurs (groupes nominaux référentiels) et par la position adoptée dans le processus de thématization

(capté dans sa phase initiale de construction discursive).

Si indexation lexicale et indexation discursive constituent une alternative, peut-on pour autant indifféremment recourir à l'une ou l'autre?

2 Indexation lexicale: éléments d'évaluation

En examinant l'évolution récente des pratiques d'indexation, on constate que indexation lexicale et indexation discursive, loin d'être en concurrence, correspondent plutôt chacune à un contexte documentaire particulier.

2.1 Notion de contexte documentaire

Par contexte documentaire, on entend l'ensemble des possibilités techniques d'accès à l'information.

De ce point de vue, on peut dégager, sur les quarante dernières années, deux contextes documentaires différents ⁽⁷⁾:

- Le contexte documentaire des années 70, où l'accès à l'information se réalise par le biais de serveurs de bases de données bibliographiques réservés à l'interrogation professionnelle. L'accès à l'information s'effectue de façon indirecte, doublement médiatisé par la référence bibliographique et par le professionnel de l'information;
- Le contexte documentaire des années 80, qui se singularise par un accès direct à l'information, c'est-à-dire un accès au «texte intégral» des

(6) «Thématiser, c'est stabiliser un état du monde raconté et se satisfaire d'un monde partiel», (Marandin 1988: 86).

(7) Il s'agit de tendance, deux contextes documentaires différents pouvant bien entendu coexister.

documents (*versus* un accès médiatisé par la référence bibliographique des documents) et un accès direct par l'utilisateur final (*versus* un accès médiatisé par les professionnels de l'information). Par exemple, l'Internet appartient à ce dernier contexte documentaire et est un exemple de ces deux traits.

Ce qui distingue radicalement le contexte documentaire actuel du contexte documentaire des années 70, c'est la perte de deux intermédiaires : l'intermédiaire textuel qu'est la référence bibliographique et l'intermédiaire humain qu'est le professionnel de l'information.

On peut formuler l'hypothèse que c'est la présence ou l'absence de ces deux intermédiaires qui impose un type d'indexation plutôt qu'un autre. C'est en reprenant la double problématique de la référence et de la thématisation que l'on peut argumenter cette position.

2.1 Adéquation de l'indexation lexicale

Dans le contexte documentaire des années 70, seule l'indexation lexicale semble possible :

- La mono-thématisation prédéfinie représente le seul moyen de rendre les documents commensurables : en l'absence du texte lui-même des documents, sur quelle autre base que la référence bibliographique (intermédiaire textuel) rassembler des documents jugés équivalents ?

Le filtrage réalisé par le langage documentaire, s'il réduit – certes – les thématiques du document à une seule, apparaît aussi comme le seul moyen d'opérer des regroupements, qui sont alors toujours prévisibles ;

- À l'époque de l'interrogation des bases de données accessibles *via* des serveurs payants, seuls les professionnels procèdent aux recherches documentaires. Eux seuls peuvent, en effet, à partir des unités

lexicales des langages documentaires, construire les référents recherchés : l'ancrage référentiel n'est pas, comme dans la langue, assuré par le groupe nominal ; il est construit *via* les relations des langages documentaires. L'intermédiaire humain est donc, dans ce cas, indispensable à l'utilisation des termes d'indexation car lui seul maîtrise les langages documentaires.

2.2 Inadéquation de l'indexation lexicale

Dès lors que le « texte intégral » devient accessible et que l'utilisateur final peut avoir directement accès aux documents, les deux intermédiaires – humain et textuel – sont perdus et l'indexation lexicale devient inadéquate :

- L'indexation « mono-thématisante » ne permet pas de donner accès au « texte intégral » d'un document : le filtrage thématique du langage documentaire devient ici contraire à l'ambition de donner accès au « texte intégral d'un document » : à quoi bon disposer de tout un texte si l'on ne peut y avoir accès que par une seule thématique, toujours la même ?

- L'utilisateur ne peut recourir aux langages documentaires pour construire les référents dont il a besoin. Les langages documentaires ne sont, pour lui, qu'un ensemble d'unités lexicales sans référence : il ne peut être sensible à l'« accord de désignation » qui les rend utilisables pour un professionnel.

Nous avons vu précédemment que la construction référentielle n'est possible, dans les langages documentaires, que si l'« accord de désignation » est premier, c'est-à-dire s'il peut supplanter l'« accord de signification » (le sens des unités lexicales). Or cette démarche va à l'encontre de la construction référentielle « naturelle » aux sujets parlants : le sens d'une unité lexicale est toujours premier et c'est toujours à

partir de lui que peut se construire, en discours, la référence (Milner 1989).

C'est pourquoi l'indexation lexicale perd de son efficacité dès lors que le texte du document est directement accessible par l'utilisateur final *via* sa fixation sur support numérique : rien ne justifie plus ni un accès partiel aux thématiques du document ni une construction référentielle médiatisée et « artificielle ».

Il faut donc envisager un autre type d'indexation.

3 Éléments de définition de l'indexation discursive

Rappelons que le double défi qui se pose actuellement aux pratiques d'indexation normatives porte sur la thématisation (qui doit être multiple) et sur la construction référentielle (qui doit être « naturelle »). L'indexation discursive va cruciallement se distinguer de l'indexation lexicale sur ces deux points.

3.1 Indexation et thématisation multiple

Nous avons vu précédemment que l'indexation discursive se situe au niveau de la construction des thèmes de discours et que c'est alors l'utilisateur qui thématise, qui achève l'interprétation d'un texte. Que fait, de son côté, l'indexeur ? L'indexeur doit maintenir le texte dans un état d'inachèvement interprétatif tout en disposant, à l'attention des utilisateurs, de quoi construire les thèmes.

Les études linguistiques menées en analyse de discours (Pêcheux 1990) ont montré que la construction thématique se réalise non pas au sein d'un seul texte mais par le biais de plusieurs textes. En effet, dans un texte, il y a des mots et ces mots n'entrent jamais « seuls » dans un énoncé : ils ouvrent au contraire « la

mémoire et l'anticipation d'autres textes» (Marandin 1984: 53); on peut dire qu'il y a toujours dans les textes des «hiatus sémantiques», des espaces où se glissent des bribes d'autres textes. Et ce sont ces «interstices», comme les appelle Foucault (1969), que l'indexeur doit restituer à l'utilisateur; en effet, comme nous l'avons vu, ce n'est pas le sens qui est en jeu dans le thème de discours mais la référence: les objets de discours d'un texte ne deviennent compréhensibles que situés, contextualisés par un ensemble d'autres textes (Foucault 1969):

«L'affirmation que la terre est ronde, et que les espèces évoluent, ne constituent pas le même énoncé avant et après Copernic, avant et après Darwin; ce n'est pas pour des formulations aussi simples que le sens des mots est changé, ce qui a été changé, c'est le rapport de ces affirmations avec d'autres propositions, ce sont leurs conditions d'utilisation et de réinvestissement, c'est le champ d'expérience, de vérifications possibles, de problèmes à résoudre auxquels on peut se référer».

Le travail de l'indexeur devient alors un travail de bâtisseur de collections: quel texte rapprocher d'un autre texte pour le rendre «lisible», exploitable, manipulable? L'indexeur se livre à une construction d'ensembles textuels qui deviennent autant de contextes permettant l'intelligibilité des textes.

L'indexation est donc recentrée désormais sur le choix des textes. Il n'est plus possible de s'en remettre aux textes tels qu'ils se donnent: il devient au contraire nécessaire d'explicitier les critères de sélection et de regroupement textuels que l'on applique⁽⁸⁾. C'est aussi de cette façon que l'indexation acquiert toute sa valeur: elle devient ce qui permet de ne pas laisser inerte et confuse la masse grandissante des textes.

L'indexation discursive, en tant que «mise en contexte des textes»,

permet en effet de manipuler le flux textuel, en établissant une «clôture» (Chartron 1997) salutaire. Grâce à cette clôture, l'Internet, par exemple, mais aussi les «flux textuels» en général, ne sont plus le tout venant du texte, ce sont des espaces organisés de documents, porteurs d'un discours spécifique⁽⁹⁾: ce que Foucault (1969: 171) propose de nommer «archive» dans son modèle:

«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'énoncés réguliers, comme autant de choses offertes au traitement et à la manipulation».

La mise en contexte des textes, opérée à la lumière des propositions de Foucault (1969), sera donc aiguillée par la question suivante: «comment se fait-il que tel énoncé soit apparu et nul autre à sa place?» (Foucault 1969: 17). Autrement dit, les regroupements textuels qui seront faits, les collections qui seront constituées, ne pourront prendre le thème comme point de repère (Foucault 1969: 16):

«Il faut renoncer à ces thèmes qui ont pour fonction de garantir l'infinie continuité du discours [...] Il faut accueillir chaque moment du discours dans son irruption d'événement; dans cette ponctualité où il apparaît, et dans cette dispersion temporelle qui lui permet d'être répété, su, oublié, transformé, effacé, jusque dans ses moindres traces [...] Il ne faut pas renvoyer le discours à la lointaine présence de l'origine; il faut le traiter dans le jeu de son instance».

Mais la création d'«espaces documentaires» (d'«archive» dans les termes de Foucault) ne suffit pas, à elle seule, pour permettre à un utilisateur de nommer ses propres thèmes.

3.2 Indexation et construction référentielle directe

L'autre caractéristique de l'indexation discursive touche la morphologie et le rôle du descripteur. Comme nous l'avons indiqué, l'indexation discursive manipule des unités de discours et des groupes nominaux en particulier. Mais nous devons ici apporter une précision: dire que la construction des thèmes de discours met en jeu non pas un texte mais plusieurs signifie que l'utilisateur dispose non pas d'unités déjà interprétées (cas des groupes nominaux d'un texte) mais d'unités interprétables dans le cadre d'une collection documentaire. En effet, le thème se laisse énoncer au terme d'une série de «saisies de différents référents discursifs» (Corblin 1995).

C'est pourquoi les descripteurs, dans l'indexation discursive, doivent être dotés de deux propriétés:

- Ils doivent permettre de parcourir une collection documentaire, *i.e.* de passer d'un texte à l'autre et de tracer un parcours interprétatif; en cela, les descripteurs doivent être pourvus de ce que l'on peut appeler une «rigidité désignative» (Corblin 1995): de ce point de vue, on dira par exemple que *école* est moins rigide que *école d'été* ou *école élémentaire*;
- Les descripteurs doivent permettre à l'utilisateur de nommer ultimement, au terme d'un parcours textuel, le thème qu'il a construit: d'un point de vue

(8) Comme le rappelle Foucault (1969: 167), il faut pouvoir dire «au nom de quoi on relie des auteurs qui ne se connaissent pas dans une trame dont ils ne sont pas maîtres?».

(9) Que l'on nommera «discours documentaire»: l'indexation comme «mise en contexte de textes» crée un nouvel espace discursif: le document, (Amar 1997)

linguistique, les descripteurs doivent donc être des unités interprétables au sein d'un groupe nominal spécifique (par exemple, *école d'été* dans *L'université organise une école d'été*).

Ces deux propriétés du descripteur se retrouvent dans un type d'unité linguistique particulier : la synapsie (David 1993).

La synapsie est un individu linguistique doté des propriétés suivantes (David 1993) :

- (i) C'est une unité de catégorie N (nom) ;
- (ii) C'est une unité qui occupe la position noyau du groupe nominal ;
- (iii) C'est une unité dont l'organisation syntaxique est constituée elle-même de trois positions ; et les positions compléments sont occupées par des compléments aréférentiels.

En tant qu'individu linguistique de catégorie N situé dans le noyau du groupe nominal (propriétés (i) et (ii)), la synapsie est une unité interprétable, et non une unité déjà interprétée.

La synapsie, créée par la syntaxe, ne correspond pas, d'un point de vue interprétatif, nécessairement à une unité lexicalisée (David 1993) : c'est cette propriété qui garantit la construction de l'interprétation, c'est-à-dire la thématization par l'utilisateur final.

La synapsie permet donc de distinguer crucialement indexation discursive et indexation lexicale, pratique qui s'assimile alors à une « manipulation de significations stabilisées », dans les termes de Pêcheux (1990 : 319-321) :

« [Il y a] une division discursive entre deux espaces : celui de la manipulation de significations stabilisées, normées par une hygiène pédagogique de la pensée, et celui des transformations du sens, échappant à toute norme assignable a priori d'un travail du sens sur le sens, pris dans la relance infinie des interprétations [...]. La conséquence de ce qui précède, c'est que toute description [...] est

intrinsèquement exposée à l'équivoque de la langue : tout énoncé est intrinsèquement susceptible de devenir autre que lui-même, de décoller discursivement de son sens pour dériver vers un autre [...]. Tout énoncé, toute séquence d'énoncés est donc linguistiquement descriptible comme une série (lexico-syntaxiquement déterminée) de points de dérive possibles, offrant place à l'interprétation ».

La synapsie est en outre une unité construite avec des compléments aréférentiels (issue de la propriété (iii)) : cette propriété lui assure une rigidité désignative.

Pour montrer le lien entre rigidité et aréférentialité des compléments, nous prendrons simplement un exemple, celui de « robe du soir » : on peut distinguer la suite *robe du soir* quand *robe du soir* est un type de robe dans un texte (par exemple, *Marie portait la robe du soir de sa mère*) de la suite *robe du soir* quand il s'agit d'une robe portée un soir (*Marie portait la robe du soir où je l'ai rencontrée*) : seul le premier *robe du soir*, où le complément « du soir » est aréférentiel (on peut porter une robe du soir le matin) est une synapsie, et seul le premier *robe du soir* intéresse l'indexation discursive dans la mesure où il peut constituer à la fois le maillon d'un parcours interprétatif et le nom d'un thème construit.

On remarque en outre que le fait que la synapsie soit interprétable, ou encore référentiellement instable, pose des problèmes de repérage par un indexeur humain : l'indexeur humain est d'abord un lecteur qui ne peut s'empêcher d'interpréter un texte : il peut difficilement dépister des unités non interprétées, c'est-à-dire situées en deçà de l'interprétation : l'indexeur humain peut donc difficilement dépister des synapsies. En effet, comment extraire une synapsie si on ne la « voit » pas, si elle ne correspond à aucun objet connu (cas des synapsies non lexicalisées, comme « robe du

matin » par exemple) ? Inversement, comment éviter d'extraire uniquement les unités que l'on connaît (cas des synapsies lexicalisées : « carte bleue » et non « carte jaune », « carton jaune » et non « carton bleu ») ?

C'est ainsi que l'on est amené, avec l'indexation discursive, à recourir – et ce nécessairement – à une extraction automatisée : le choix des mots est désormais, en indexation, confié à un programme informatique, seul capable d'identifier ce type d'unités situées en deçà de l'interprétation. Comment désormais va se distribuer le travail entre la machine et l'indexeur ?

3.3 Vers une pratique de l'indexation discursive

L'indexation discursive se laisse donc, dans ses grandes lignes, décrire ainsi :

- Elle est d'abord et avant tout une opération de constitution de collections, de mise en contexte des textes : on peut, sur ce premier aspect, prendre pour point de départ les propositions de Foucault (1969). L'essentiel du travail de l'indexeur se déplace donc du lexique (travail sur les langages documentaires) au discours (travail sur la constitution de collections, devenant « discours documentaires », (Amar 1997)) ;
- Elle est ensuite une opération d'extraction automatisée de synapsies : c'est ici nécessairement la machine qui identifie les unités permettant la construction des parcours discursifs et la nomination des thèmes : sur ce deuxième aspect, de quel programme informatique dispose-t-on ? Quelle méthode peut utiliser l'indexeur pour sélectionner les synapsies qui permettront au lecteur de construire ses thèmes ?

Certes, on compte peu de logiciels susceptibles de réaliser aujourd'hui une identification des synapsies d'un texte (David 1993 :

chapitre V). Sur ce point, il importe de signaler le travail réalisé dans les années 1990 sur le logiciel *Termino* ⁽¹⁰⁾ (David et Plante 1990 et 1991, David 1993) : ce logiciel, fondé uniquement sur un savoir linguistique, permet, grâce à la description syntaxique des textes qu'il met en œuvre, de proposer à l'utilisateur du logiciel la liste des synapsies qu'ils comprennent ⁽¹¹⁾. *Termino* donne aux pratiques de l'indexation discursive une possibilité de prendre corps : il constitue une condition nécessaire à l'indexation discursive, mais pas une condition suffisante.

L'indexeur prend ensuite le relais de la machine pour sélectionner les synapsies à proposer à l'utilisateur d'un système documentaire. Il doit retenir des unités permettant de rendre compte des choix qu'il a effectués en amont sur les textes et leur organisation. L'indexeur construit, comme nous l'avons vu, une collection documentaire présentant une certaine «cohérence discursive»; cette cohérence est ensuite rendue accessible, visible, palpable par les descripteurs mais ce ne sont pas eux qui déterminent les filiations, les accointances entre textes : ils les montrent simplement (rôle de l'index précisément). Dans l'indexation discursive, les mots sont l'effet, le résultat visible et palpable d'un processus antérieur de regroupement textuel, ils n'en sont pas la cause.

Nous proposons, à titre indicatif dans cette contribution ⁽¹²⁾, les «chaînes de référence» comme point de départ à la constitution d'une méthode de sélection des synapsies.

La notion de chaîne de référence, proposée par Chastain (1975) et travaillée notamment par Marandin (1988) et Corblin (1995), indique comment une unité comme la synapsie peut être utilisée pour créer des parcours textuels dans une collection de documents, pour permettre la «transformation du sens» et l'exploitation de l'«équivoque de la

langue», pour reprendre la problématique de Pêcheux (1990) précédemment abordée.

Une chaîne de référence peut être entendue (Corblin 1995 : 27) comme «la suite d'expressions [...] entre lesquelles l'interprétation établit une identité de référence». L'intérêt de la notion de chaîne de référence est qu'elle prend en compte la diversité formelle qui permet à l'interprétation (à la référence) de se construire (Corblin 1995 : 174) :

«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».

La notion de chaîne de référence, ici très brièvement évoquée, montre que le parcours interprétatif dans une collection de documents ne saurait se réduire à la seule possibilité technique de navigation hypertextuelle : pour que ce parcours soit effectif (*i.e.* pour que le lecteur parvienne réellement à construire des thématiques), il est indispensable de lui proposer des points de repère linguistiquement circonscrits, appartenant à des chaînes de référence, c'est-à-dire susceptibles de capter - au travers d'une diversité lexicale, syntaxique et discursive - une «identité référentielle», un thème possible, les traces d'un objet en

construction. L'interprétation thématique, si elle est à la fois multiple et mouvante, n'est pas pour autant - et loin s'en faut - anarchique : elle reste toujours gouvernée par des règles linguistiques et discursives. Ces règles, déjà respectées lors de la mise en contexte des textes, sont à nouveau exploitées pour établir le cheminement à travers textes.

À travers cette ébauche de ce que peut être l'indexation discursive, n'est-on pas amené à se réinterroger sur la spécificité même de l'indexation ? Certes, comme nous l'avons dit, l'indexation discursive est particulièrement adéquate dans le contexte documentaire actuel, mais ne l'est-elle pas aussi au-delà ?

Les possibilités techniques actuelles, en permettant au «texte intégral» des documents d'être directement l'objet de traitement documentaire, sans médiation nécessaire, nous donnent la possibilité de renouer avec l'objet premier de l'indexation - les textes ⁽¹³⁾ - et avec sa caractéristique première - la manipulation de textes. Du coup

(10) *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.

(11) Signalons que le logiciel *Termino* ne traite, dans ses versions actuelles, que de textes écrits en français.

(12) On peut se reporter à Amar (1997) pour plus de détails sur l'utilisation des chaînes de référence dans un contexte documentaire.

(13) Certes, tous les objets de l'indexation ne sont pas des objets textuels : la nature sémiotique des objets d'indexation non-textuels ne doit pas cependant être négligée : sur ce point, on peut se reporter à (Amar 1997 : chapitre 1).

n'est-on pas conduit à redéfinir l'indexation, en toute généralité, comme un traitement textuel, c'est-à-dire un traitement permettant de saisir la spécificité sémiotique des objets qu'elle manipule? Dans ce cas, l'indexation lexicale pourrait n'apparaître que comme un accident historique qui, sous le poids des contraintes techniques, aurait perdu de vue la finalité de l'indexation, rendre compte des textes?

Conclusion

En exploitant l'opposition lexicale/discours établie en linguistique, on peut montrer qu'il existe deux types d'indexation – lexicale et discursive –, radicalement différents dans leur approche des textes comme dans les unités qui leur donnent accès. Alors que le domaine d'exercice de l'indexation lexicale semble limité à un contexte documentaire spécifique, l'indexation discursive, qui se donne les moyens de rendre compte de la spécificité des objets qu'elle manipule, apparaît plus largement applicable. Reste que de nombreux aspects de l'indexation discursive doivent être approfondis avant de pouvoir recouvrir tous les faits d'indexation, en toute généralité: notamment les critères de sélection et de regroupement des textes et la méthode de sélection des synapsies, pour lesquels on a proposé ici que des pistes.

D'ores et déjà, il apparaît que l'indexation normative mérite d'être «historiquement» évaluée. C'est donc indirectement la fonction «normative» des textes professionnels qui est ici interrogée: quelle est la durée de validité d'une norme? Quel statut épistémologique peut-on lui attribuer? Sur quoi fonder les pratiques professionnelles, norme ou théorie?

Notre essai de typologie de l'indexation (lexicale/discursive) permet en outre de disposer

d'éléments pour évaluer des procédures d'indexation automatique qui, tout en exploitant des unités de discours (le plus souvent les groupes nominaux des textes), réinvestissent le champ lexical *via* des réseaux sémantiques ou encore des thésaurus (comme *Worldnet*, Habert *et al.* 1997): de quel type d'indexation relèvent ces pratiques, lexicale ou discursive? Est-ce parce que les procédures d'indexation automatique traitent le «texte intégral» des documents qu'elles en effectuent, pour autant, le traitement textuel?

Muriel Amar,
*Langages, cognitions,
pratiques et ergonomie,
Institut national de la langue française,
Paris.*

Bibliographie

Afnor 1978: «Norme Z 47-102: principes généraux pour l'indexation des documents» dans *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» dans *Documentation, Tome 1: présentation des publications, traitement documentaire et gestion des bibliothèques*, Paris, Afnor (Recueil de normes françaises).

Amar (M.) 1997: *Les fondements théoriques de l'indexation: une approche linguistique*, Doctorat en sciences de l'information et de la communication, Lyon, Université Lumière-Lyon II. À paraître aux éditions ADBS en mars 2000.

Barthes (R.) 1970: *S/Z*, Paris, Seuil (Tel Quel).

Chartron (G.) 1997: *Repérage de l'information sur Internet: nouveaux outils, approches bibliothéconomiques et micro-structures*, [texte de la communication donnée lors du] Stage CRDP de Reims *CDI virtuel, utilisation des réseaux pour l'accès à la communication*.

Corblin (Fr.) 1995: *Les Formes de reprise dans le discours: anaphores et chaînes de référence*, Rennes, Presses universitaires de Rennes.

David (S.), Faraj (N.), Godin (R.), Missaoui (R.) et Plante (P.) 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, pp. 1-21.

David (S.) 1993: *Les unités polylexicales: éléments de description et reconnaissance automatique*, Thèse de doctorat en linguistique théorique et formelle, Paris, Université de Paris VII.

David (S.) et Plante (P.) 1990: «De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes», *ICO*, vol. 2, n° 3, pp. 140-155.

David (S.) et Plante (P.) 1991: «*Termino* v. 1.0TM: principes et propriétés linguistiques» dans *Actes du colloque Industries de la langue*, novembre 1990, Montréal, OLF et Société des traducteurs du Québec, pp. 71-88.

Dictionnaire de linguistique et des sciences du langage, 1994: Paris, Larousse.

Dubois (D.) et Mondada (L.) 1995: «Construction des objets de discours et catégorisation: une approche des processus de référenciation», *Tranel*, n° 23, pp. 273-302.

Foucault (M.) 1969: *Archéologie du savoir*, Paris, Gallimard.

Habert (B.), Nazarenko (A.) et Salem (A.) 1997: *Les linguistiques de corpus*, Paris, A. Colin (U. Série Linguistique).

Jacquesson (A.) et Rivier (A.) 1999: *Bibliothèques et documents numériques: concepts, composantes et enjeux*, Paris, Éditions du Cercle de la librairie (Bibliothèques).

Le Guern (M.) 1984: «Les Descripteurs d'un système documentaire: essai de définition», *Condenser*, Suppl. 1, pp. 163-169.

Le Guern (M.) 1989: «Sur les relations entre terminologie et lexique», *Méta*, vol. XXXIV, n° 3, pp. 340-343.

Le Guern (M.) 1991: «Un Analyseur morpho-syntaxique pour l'indexation

automatique», *Le français moderne*,
1 (59), pp. 22-35.

Marandin (J.-M.) 1984: «Mais qu'est-ce
que Socrate a au juste à voir avec la
sagesse?», dans *Linx*, n° 10, pp. 51-55.

Marandin (J.-M.) 1988: « À propos de la
notion de thème de discours. Éléments
d'analyse dans le récit », *Langue française*,
n° 78, pp. 67-87.

Marandin (J.-M.) 1997: *Perception
syntaxique et constructions syntaxiques*,
Mémoire d'habilitation, Paris, Université
Paris VII-Denis Diderot.

Milner (J.-Cl.) 1989: *Introduction à une
science du langage*, Paris, Seuil.

Pêcheux (M.) 1990: *L'inquiétude du
discours*, [Textes choisis et présentés par
Denise Maldidier], Paris, Éditions des
Cendres.

Simonot (Fr.) 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.

Thésaurus de l'Unesco, 1995: Paris,
Unesco.