

# Indexation manuelle et indexation automatique. Dépasser les oppositions

*Auteurs : Ghislaine CHARTRON\*, Sylvie DALBIN\*\* Marie-Gaëlle MONTEIL \*\*, Monique VERILLON\*\**

Documentaliste, vol. 26, n° 4-5, juillet-octobre 1989, p.181-187

Mis en forme en 2006 par Sylvie Dalbin (sylvieatd@aol.com)

=====  
Résumé

La présente étude rend compte d'un travail de comparaison entre l'indexation manuelle et l'indexation automatique d'un même corpus de documents. La première a été faite à l'aide du thésaurus EDF, la seconde à l'aide du système LEX/NET dont l'organisation et le fonctionnement sont présentés. Le corpus considéré rassemble près de 2 400 notices de la base EDF-DOC relatives à un unique sujet: l'intelligence artificielle. Après avoir comparé les résultats obtenus avec les deux méthodes, l'analyse met en évidence les traits dominants, les insuffisances et les avantages de chacune. Enfin, la qualité des indexations obtenues est évaluée au regard de quelques critères: pertinence, accessibilité, cohérence et évolutivité. Mais la meilleure indexation ne serait-elle pas celle qui allierait les deux méthodes.

- =====  
  - [1. Contexte des travaux](#)
  - [2. Le système Lexinet](#)
  - [3. L'indexation manuelle](#)
  - [4. Chiffres et courbes des deux indexations](#)
  - [5. Analyse qualitative des deux indexations](#)
    - [5.1. L'indexation manuelle](#)
    - [5.2. L'indexation Lexinet](#)
  - [6. Synthèse des caractéristiques d'indexation](#)
    - [6.1. Description pertinente](#)
    - [6.2. Accessibilité maximale](#)
    - [6.3. Cohérence d'indexation](#)
    - [6.4. Evolutivité](#)
  - [7. Faut-il associer les deux modes d'indexation ?](#)
  - [Bibliographie](#)

## 1. Contexte des travaux

Ce travail de comparaison d'indexations, mené conjointement par une équipe de l'INIST/CNRS et le Centre de documentation de la Direction des études et recherches d'Electricité de France, s'inscrit dans un contexte déjà ancien d'expérimentations de systèmes documentaires informatisés, généralement basés sur des méthodes linguistiques, statistiques, combinatoires ou relevant de l'intelligence artificielle (systèmes SPIRIT, DIALECT...).

Le système expérimenté pour l'indexation automatique, LEXINET, est prioritairement un gestionnaire de lexiques. Fondé sur des procédés statistiques et combinatoires, il permet de construire ces lexiques dans les domaines scientifiques et techniques à partir du vocabulaire repéré dans les corpus de documents. La performance d'un tel système n'est envisagée que sur des textes condensés, type résumé d'article. LEXINET présente certaines caractéristiques communes avec les systèmes SMART [8], SPIRIT [4] et le projet TINA [10], et également avec les travaux menés en statistique textuelle [7]. De plus, ce système permet d'obtenir une indexation de chacun des documents en utilisant le lexique construit par apprentissage sur les textes, cet

apprentissage étant guidé par un expert humain, spécialiste du domaine traité.

Au cours des essais de ce système, étudié a priori pour constituer ou mettre à jour un vocabulaire de spécialité, nous avons été amenés à comparer les deux indexations, indexation manuelle pratiquée par EDF et indexation Lexinet, et à relever les caractéristiques, défauts et avantages de l'une et l'autre obtenues sur un même corpus de documents. Ce corpus étudié rassemble 2 380 résumés et titres de la base EDF-DOC sur le domaine de l'intelligence artificielle. Le choix de ce domaine n'était pas dû au hasard: les problèmes de mise à jour du thésaurus, dans un domaine aussi évolutif, créaient un décalage important du vocabulaire existant dans le thésaurus par rapport à la réalité

## 2. Le système Lexinet

Lexinet est un système performant sur un corpus de textes scientifiques et techniques condensés (résumés de brevets, de publications scientifiques) d'un domaine, précis, qui pourra être aussi bien la biotechnologie que les céramiques électriques ou l'intelligence artificielle. Sur les serveurs, les documents circulent sous une forme «titre et résumés », les techniques développées trouvant ainsi un vaste champ d'applications.

Compte tenu de la syntaxe simplifiée des données textuelles à traiter, le choix effectué pour le développement de Lexinet a été d'utiliser des procédés statistiques et combinatoires faciles à mettre en œuvre et répondant aux exigences d'opérationnalité du système dans un domaine quelconque. Des algorithmes flexibles ont été réalisés pour effectuer des traitements linguistiques élémentaires : normalisation des termes, détection des mots composés, synonymie des termes. Certains de ces algorithmes rejoignent des méthodes préconisées par Salton [9].

Les traitements Lexinet ont été envisagés comme un ensemble de modules, chacun de ceux-ci effectuant une tâche spécifique. Les principaux modules développés sont les suivants :

- découpage des textes : le traitement consiste à repérer chaque mot (uniterme) avec toutes ses localisations dans le corpus étudiés ;
- normalisation des termes : on effectue un regroupement des termes selon une étude des suffixes ;
- filtrage de mots vides : on se réfère à une liste de mots vides prédéfinis ;
- filtrage distributionnel : le traitement consiste à éliminer les termes usuels de la langue en étudiant leur distribution dans le corpus de textes traité ;
- gestion des synonymies : on utilise un dictionnaire de synonymes construit par apprentissage ;
- reconnaissance et gestion des termes composés : les termes composés sont repérés par une étude statistique de proximité de chaînes de caractères ;
- gestion des unitermes ;
- fonction d'indexation des documents.

Ces différents modules s'articulent dans un scénario de traitement où l'expert humain est sollicité pour effectuer des choix finals. Il intervient alors interactivement à l'écran et corrige certains résultats proposés par l'ordinateur.

Trois unités de données sont continuellement et automatiquement mises à jour lors de l'enchaînement des traitements : ce sont le lexique, l'antilexique et le dictionnaire de synonymies. Le lexique contient tous les termes retenus car significatifs ; l'antilexique contient tous les termes rejetés ; le dictionnaire de synonymies stocke tous les liens de report établis entre les termes. Au départ, les dictionnaires peuvent être vides ; ils se construisent au fur et à mesure d'un apprentissage sur les textes. Les allers-retours entre les procédures automatisées et le jugement d'expert humain aboutissent finalement à la création d'un lexique des termes significatifs du corpus de documents traité. Ce lexique peut être corrigé par différentes procédures interactives.

### 3. L'indexation manuelle

L'indexation manuelle, faite par des spécialistes des techniques abordées dans les documents, s'effectue à partir du thésaurus EDF. Dans les domaines en forte évolution – c'est le cas de l'intelligence artificielle – il existe un décalage important du thésaurus par rapport à la technique, décalage compensé par l'utilisation de « candidats descripteurs ». L'indexation manuelle du corpus étudié inclut ces candidats descripteurs.

Cette indexation manuelle se base sur quatre points essentiels :

- la lecture du document dans son intégralité pour l'élaboration de l'indexation ;
- la prise en compte, dans le choix des descripteurs, des objectifs du centre de documentation (applications) et des besoins des utilisateurs ;
- la complémentarité permanente entre les termes de l'indexation manuelle et le résumé;
- en l'absence de descripteur approprié, et lorsque l'émergence d'un nouveau concept n'est pas suffisamment explicite pour proposer un candidat descripteur, la possibilité d'utiliser un descripteur voisin ou générique

### 4. Chiffres et courbes des deux indexations

L'application de Lexinet sur le corpus a nécessité deux semaines de travail avec les documentalistes d'EDF; nous avons obtenu finalement à partir du lexique constitué lors de la première phase une « indexation Lexinet » pour chaque document, qui a pu être comparée à l'indexation manuelle existante.

Si nous opérons un classement des termes d'indexation par occurrences décroissantes, nous pouvons affecter un rang à chaque terme et lui associer son occurrence respective.

Nous avons intitulé « Classement des termes d'indexation » les courbes ainsi obtenues (fig. 1). La courbe « indexation manuelle » se positionne nettement au-dessus de la courbe « indexation Lexinet » jusqu'au 20e rang. L'écart tend ensuite à diminuer et une représentation complète montre qu'au 1 500e rang la courbe « indexation Lexinet » est passée au-dessus, le mot 1 500 ayant une occurrence de 3 pour Lexinet, et de 1 pour l'indexation manuelle.

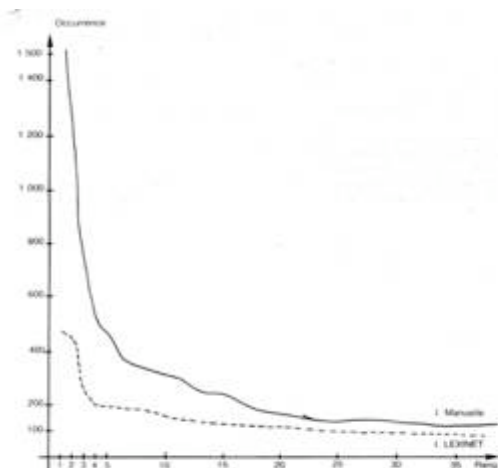


Figure 1. – Classement des termes d'indexation

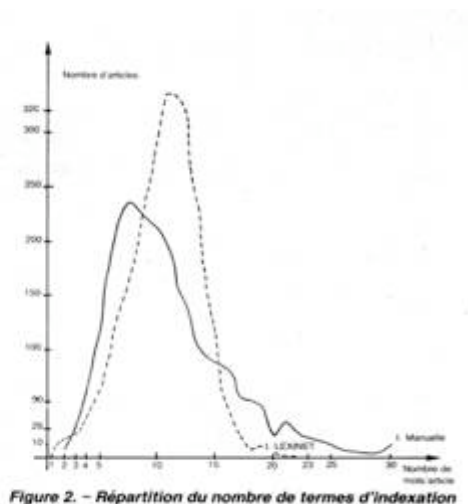


Figure 2. – Répartition du nombre de termes d'indexation

Les courbes montrent que l'indexation manuelle utilise un vocabulaire plus restreint; les très fortes fréquences des premiers mots traduisent une utilisation systématique de « macrotermes » (cf. tableau 1). La décroissance de la courbe Lexinet montre l'utilisation d'une plus grande variété de termes traduisant simultanément un effet de dispersion et un effet de spécificité du vocabulaire.

Mot	Fréquence indexation manuelle	Fréquence Lexinet
intelligence artificielle	1523	410
reconnaissance des formes	1012	251
système expert	699	435
algorithme	505	189
traitement d'image	417	115
modélisation	367	091
programmation	348	081
logiciel	340	134

Tableau 1 : Termes les plus fréquents pour l'indexation manuelle, fréquences comparées à celles de Lexinet.

En effectuant une moyenne du nombre de termes indexant chaque article, nous aboutissons aux résultats suivants:

- pour l'indexation manuelle moyenne = 10,89
- pour l'indexation Lexinet moyenne = 10,36

Nous aboutissons à deux moyennes similaires. Cependant l'étude de la distribution du nombre d'articles associés à un certain nombre de mots d'indexation montre une répartition différente (cf. en fig. 2, les courbes « Répartition du nombre des termes d'indexation »). La courbe relative à Lexinet présente un pic marqué pour une valeur voisine de 10 mots d'indexation. La courbe relative à l'indexation manuelle présente un sommet moins marqué (pour une valeur de 7 mots d'indexation) ; elle tend à s'étaler sur une plage assez large du nombre de mots; nous observons des variations importantes marquées de petits pics. L'écart à la valeur moyenne est plus important pour la courbe indexation manuelle que pour la courbe Lexinet, ce qui semble traduire une régularité plus grande dans l'indexation Lexinet. Le nombre de descripteurs semble être plus variable pour l'indexation manuelle, plusieurs paramètres semblant intervenir:

- une indexation différente selon le type de documents: articles, rapports
- de congrès, livres...
- l'absence de certains concepts dans le thésaurus,
- les méthodes différentes de travail des indexeurs.

## 5. Analyse qualitative des deux indexations

Entre les deux modes d'indexation, il faut noter déjà un effet de contenu du document : l'indexation lexinet opère seulement sur le titre et le résumé du document alors que pour l'indexation manuelle, l'indexeur dispose de l'intégralité du document et peut envisager une complémentarité entre indexation et résumé. Ceci explique que l'indexation du rédacteur nous soit apparue parfois décalée par rapport au résumé. Cet effet de contenu provient donc du fait que les deux indexations ne sont pas construites sur les mêmes entités textuelles.

### 5.1. L'indexation manuelle

Plusieurs traits caractérisent l'indexation manuelle.

**1. Un effet de généralisation** lié à un effet de synthétisation de l'information. Certains termes généraux sont introduits de façon quasi systématique par les indexeurs dans le but de « traduire » le contenu des documents pour des demandeurs potentiels travaillant dans des disciplines variées. Les fortes occurrences des termes *intelligence artificielle*, *reconnaissance des formes*, *système expert* révèlent une habitude à indexer par ces termes même si un degré plus fin de spécificité peut être atteint. *Reconnaissance des formes* est employé pour la reconnaissance vocale, la reconnaissance des caractères et de l'image. L'indexation par ces termes généraux est une façon de créer une structure analogue à celle d'un plan de classement.

**2. Une amplification** de certains éléments des documents en rapport avec une volonté de mettre

en valeur les applications mentionnées dans les documents. Cette amplification des applications est relative aux orientations que se fixe le centre documentaire, orientations liées aux demandes des utilisateurs de la base de documents. Cette remarque se vérifie surtout dans le cas d'une base d'entreprise, comme la base EDF-Doc dont les documentalistes connaissent majoritairement les utilisateurs.

Dans le document reproduit dans le Tableau 2, l'indexation manuelle met en valeur les applications potentielles par les descripteurs gestion financière, médecine, politique d'investissement. L'indexation Lexinet référence les applications de cardiologie, pathologie, biomédecine que l'indexation manuelle compacte par le terme médecine. Elle ne rend pas compte de l'application aux problèmes financiers (pas de terme retenu), alors que l'indexation manuelle souligne cette application par les termes: gestion financière et politique d'investissement.

<p>TITRE : RECONNAISSANCE DES FORMES          RESUMÉ : - DESCRIPTION D'UN SYSTÈME DE PROGRAMMATION EXTENSIBLE ET CONVERSATIONNEL DESTINÉ A FACILITER L'UTILISATION DES CALCULATEURS POUR LES PROBLÈMES DE LA BIOMÉDECINE (CARDIOLOGIE, PATHOLOGIE).          - DESCRIPTION DE LA CONCEPTION ET DES PRINCIPES D'IMPLANTATION D'UN SYSTÈME CONVERSATIONNEL D'ANALYSE ET DE CLASSIFICATION DE FORMES SUR CONSOLE GRAPHIQUE.          - RÉSUMÉ DE QUELQUES PROJETS RÉCENTS APPLIQUANT LE TRAITEMENT NUMÉRIQUE DES IMAGES A LA GESTION DES RESSOURCES NATURELLES.          - PRÉSENTATION DES RÉSULTATS D'UN PROGRAMME DE RECHERCHE SUR LES APPLICATIONS DES CONCEPTS CYBERNÉTIQUES ET DES TECHNIQUES DE L'INTELLIGENCE ARTIFICIELLE DANS LA RÉOLUTION AUTOMATISÉE DES PROBLÈMES FINANCIERS (INVESTISSEMENTS).</p>	
Indexation Lexinet :	Indexation manuelle
1.00000 RECONNAISSANCE FORME	<u>GESTION FINANCIERE</u>
0.52500 <u>CARDIOLOGIE</u>	<u>IEEE</u>
0.27500 <u>PATHOLOGIE</u>	<u>INTELLIGENCE ARTIFICIELLE</u>
0.19167 RESSOURCE NATURELLE	<u>MEDECINE</u>
0.15000 BIOMEDECINE	<u>MODE CONVERSATIONNEL</u>
0.15000 CYBERNETIQUE	<u>POLITIQUE D'INVESTISSEMENT</u>
0.07500 CONSOLE GRAPHIQUE	<u>RECONNAISSANCE DES FORMES</u>
0.06071 CLASSIFICATION FORME	<u>SORTIE GRAPHIQUE</u>
0.05441 PROGRAMME RECHERCHE	<u>WASHINGTON</u>
0.04583 IMPLANTATION	1975
0.04352 TRAITEMENT NUMERIQUE IMAGE	
0.03816 RESUME	
0.03194 TRAITEMENT NUMERIQUE	
0.03185 RESOLUTION	
<p>(Les coefficients figurant devant chaque terme d'indexation Lexinet correspondent au poids affecté à chaque terme par le calcul d'une fonction statistique mesurant la représentation du terme avec le document et inversement.)</p>	

Tableau 2

**3. Un effet de variabilité** : nous avons déjà évoqué le fait que deux indexeurs choisissent rarement exactement les mêmes termes pour indexer un même document. On a observé, parfois, l'absence de certains descripteurs que l'indexeur aurait dû retenir pour rendre compte du contenu.

**4. Un effet de décalage** entre le thésaurus et le contenu des documents, provenant des délais inévitables de mise à jour manu,elle d'un thésaurus. Cet effet est accentué dans des domaines qui évoluent très rapidement tels que l'intelligence artificielle. Cet effet de décalage est aussi lié à l'utilisation même d'un langage contrôlé.

## 5.2. L'indexation Lexinet

Plusieurs caractéristiques principales se dégagent de l'indexation Lexinet.

Un effet d'ambiguïté de certains termes pris en dehors de leur contexte. Prenons l'exemple du terme *synthèse* identifié dans les deux documents suivants (tableaux 3 et 4).

<p>TITRE : ETUDE DES CLÉS D'IDENTIFICATION ET DES TABLES DE DIAGNOSTIC          RESUMÉ : - ÉTUDE DE LA MÉTHODOLOGIE ET DES DOMAINES D'APPLICATION DES CLÉS D'IDENTIFICATION ET DES TABLES DE DIAGNOSTIC, EN TENANT COMPTE DE LA THÉORIE MATHÉMATIQUE ET DES SPÉCIFICATIONS PRATIQUES. DES TECHNIQUES PROBABILISTES ET NON-PROBABILISTES SONT ÉTUDIÉES. L'ARTICLE ESSAYE DE FAIRE UNE <u>SYNTHÈSE</u> D'UNE IMPORTANTE LITTÉRATURE LARGEMENT DISPERSÉE QUI N'EST PAS CONVENABLEMENT INTÉGRÉE. EN CE SENS QUE DES CHERCHEURS DANS UN DOMAINE D'APPLICATION IGNORAIENT SOUVENT LA THÉORIE ASSOCIÉE DÉVELOPPÉE DANS D'AUTRES DOMAINES.</p>	
Indexation Lexinet :	Indexation manuelle
1.00000 CLE	<u>DOCUMENT DE SYNTHÈSE</u>
1.00000 DIAGNOSTIC	IDENTIFICATION (SYSTEME)
1.00000 IDENTIFICATION	DIAGNOSTIC
1.00000 TABLE	MEDECINE
0.19762 THEORIE MATHÉMATIQUE	REPRESENTATION (SYSTEME)
0.11513 PROBABILISTE	STRUCTURE ARBORESCENTE
0.06250 LITTÉRATURE	DETECTION DE DEFAULT
0.03750 SPECIFICATION	RECONNAISSANCE DES FORMES
0.03713 METHODOLOGIE	THEORIE DU CODAGE
0.03510 <u>SYNTHÈSE</u>	TABLE DE DECISION
0.03379 THEORIE	THEORIE DES PROBABILITES
	TEST SEQUENTIEL
	OPTIMISATION
	ALGORITHME
	THEORIE DE LA DECISION

**Tableau 3**

<p>TITRE : <u>SYNTHÈSE</u>, RECONNAISSANCE DE LA PAROLE          RESUMÉ : PREMIER LIVRE S'ADRESSANT DIRECTEMENT AUX ROBOTS QUI SOUHAITENT APPRENDRE A PARLER OU A ECOUTER LEUR MAITRE. CHAPITRE 1 : COMMENT EN EST-ON ARRIVE LA ? CHAPITRE 2 : A QUOI SERT-IL DE PARLER ? CHAPITRE 3 : COMMENT CELA MARCHE ? (LE CONDUIT VOCAL - L'ANALYSE - LE CODAGE - LA <u>SYNTHÈSE</u> - LA RECONNAISSANCE). CHAPITRE 4 : LES CIRCUITS DE PAROLE DU MARCHE. CHAPITRE 5 : LES CIRCUITS...</p>	
Indexation Lexinet :	Indexation manuelle
1.00000 <u>SYNTHÈSE</u>	INTELLIGENCE ARTIFICIELLE
1.00000 RECONNAISSANCE PAROLE	LANGUE (ORGANE)
0.22222 CONDUIT VOCAL	CIRCUIT ELECTRONIQUE
0.06349 LIVRE	<u>SYNTHÈSE (SYSTEME)</u>
0.05913 CODAGE	AUTOMATISME
0.05905 PAROLE	INFORMATIQUE
0.05790 ROBOT	BUREAUTIQUE
0.05730 RECONNAISSANCE	TELEINFORMATIQUE
	LARYNX
	MATERIEL INFORMATIQUE (REFERENCE)

**Tableau 4**

Dans le premier document, *synthèse* est utilisé dans le sens document de synthèse. Dans le deuxième document *synthèse* représente une technique, précise en intelligence artificielle, à savoir la synthèse de la parole. Avec Lexinet, on a indexé les deux documents par le terme *synthèse*, confondant les deux significations. L'indexation manuelle lève cette ambiguïté en

indexant le premier document par document de synthèse et le deuxième par synthèse (système). Remarquons aussi que dans le premier document, le terme clé est également un terme dont la signification risque d'être ambiguë. Dans ce document, clé fait référence à une clé de codage dans un contexte mathématique. En revanche, on trouve d'autres documents indexés par clé, mais employé alors avec une autre signification et ne s'assimilant pas à un terme significatif du contenu du document. Par exemple on trouve le contexte suivant : "la clés du succès dépend d'une utilisation...".

**1. Un effet de dispersion de l'information** lié à une absence de termes synthétiques permettant de recouvrir une certaine catégorie d'information comme le font certains termes de l'indexation manuelle. Nous avons déjà cité l'exemple du descripteur *reconnaissance des formes* en mentionnant l'effet opposé de généralisation de l'indexation manuelle.

**2. Une certaine cohésion d'indexation.** La sélection d'un terme significatif pour le corpus implique sa projection sur l'indexation de tous les documents qui le contiennent et pour lesquels son coefficient d'indexation est supérieur au seuil fixé. Cette façon de procéder assure une cohésion dans la pratique d'indexation.

**3. Une description spécifique de l'information,** opposée à l'effet de généralisation mais provoquant en partie l'effet de dispersion.

## 6. Synthèse des caractéristiques d'indexation

Que demande-t-on à un système sommes restés dans ce cadre d'indexation ?

L'utilisation finale de l'indexation peut conditionner des exigences différentes. On évalue principalement une indexation par rapport à sa performance pour la recherche rétrospective des documents [9,12]. Mais l'indexation d'un corpus de documents peut servir d'autres objectifs que celui-ci; elle peut, en particulier, être utilisée pour élaborer des vues synthétiques des informations stockées dans un corpus [2]. On peut alors envisager d'autres modes d'évaluation de l'indexation [11].

Mais quel que soit le contexte d'utilisation de l'indexation, l'objectif fondamental de cette opération est le signalement optimal du contenu des documents. Pour caractériser les deux indexations obtenues, nous sommes restés dans ce cadre très général du rôle de l'indexation pour définir les critères d'appréciation correspondants [1,5].

### 6.1. Description pertinente

L'indexation doit offrir une description pertinente. La pertinence peut se mesurer essentiellement par l'exhaustivité, la spécificité, le degré d'ambiguïté des termes.

**1. L'exhaustivité** caractérise la capacité à signaler toutes les notions importantes d'un document. Nous pouvons différencier d'une part une exhaustivité par rapport au domaine traité, ce que l'on a appelé une *exhaustivité interne au domaine*, et d'autre part une exhaustivité par rapport à des notions connexes évoquées dans les documents mais non spécifiques du domaine traité, ce que l'on a appelé une *exhaustivité externe*.

Pour l'indexation manuelle, l'effet de décalage entre le thésaurus et les contenus des documents (observé dans le cas précis d'un domaine en pointe tel que l'intelligence artificielle), ainsi que l'effet de variabilité du choix de descripteurs sont parfois les causes d'une exhaustivité interne insuffisante. En revanche, nous trouvons une bonne exhaustivité externe par rapport aux termes relatifs aux applications évoquées dans les documents, ce que nous avons souligné par un effet d'amplification dans nos observations précédentes.

Pour l'indexation Lexinet, l'exhaustivité dépend étroitement des choix effectués lors des listes présentées à l'expert pour constituer le lexique initial mais aussi de la qualité des résumés. Dans cette étude, l'objectif était de constituer un vocabulaire représentatif de l'intelligence artificielle. En conséquence, on a obtenu une bonne exhaustivité interne avec une prise en compte d'un vocabulaire

récent tel que *langage orienté objet, administrateur de base de données, base de règles*.

Par contre, étant donné l'objectif fixé, les experts n'ont pas cherché à garder particulièrement des notions connexes au domaine traité. Les experts ont pu rejeter certains termes qui, dans un objectif différent, auraient pu être gardés.

Dans un cas de pauvreté du résumé (et du titre), la qualité de l'indexation Lexinet s'en ressentira immédiatement, au risque de créer du « silence » lors de la recherche. Alors que dans le cas d'un thésaurus pauvre, l'indexeur peut toujours modifier et enrichir le résumé pour que la recherche (en langage libre, complémentaire à la recherche en langage contrôlé) soit performante.

**2. La spécificité** caractérise la capacité à garder une information précise et non généralisée. L'indexation Lexinet rend compte d'un niveau supérieur de spécificité ; on rend compte de l'information telle qu'elle est citée dans les textes, sans effet de généralisation. En moyenne, l'indexation manuelle est moins spécifique mais plus synthétique.

**3. Le degré d'ambiguïté** représente la polysémie possible des termes d'indexation. Nous avons vu précédemment que l'indexation Lexinet est peu performante sur ce point. Par contre, pour l'indexation manuelle, le problème d'ambiguïté ne se pose pas car les descripteurs du thésaurus sont répartis en 320 « champs sémantiques » et chaque descripteur n'est associé qu'à un seul champ. Ainsi le descripteur *synthèse*, par exemple, ne peut être utilisé que dans le sens d'une synthèse effectuée par un système automatique et non dans le sens d'une opération mentale résumant un ensemble d'informations. Si l'on veut employer ce terme dans ce deuxième contexte, il faudra utiliser un autre descripteur prévu dans ce cas précis, à savoir le descripteur document de synthèse.

Ces propriétés de l'indexation (exhaustivité, spécificité, degré d'ambiguïté) nous semblent nécessaires pour alimenter aussi bien des systèmes d'interrogation que des systèmes visant à donner des vues synthétisées d'un ensemble de documents en utilisant les mots-clés représentatifs de cet ensemble. On peut tout de même nuancer la nécessité de lever l'ambiguïté qui est primordiale pour une interrogation optimale dès documents mais moins gênante dans une synthèse de contenu global où l'indexation isolée de quelques documents par un terme inadéquat disparaît dans les statistiques générales que l'on peut établir.

## 6.2. Accessibilité maximale

Un système d'indexation doit permettre une accessibilité optimale de l'information. Des catégories intermédiaires de mots sont utiles pour permettre au demandeur d'informations de disposer de différents niveaux de description des informations, ce qui lui permet généralement d'affiner progressivement sa requête. Ces mots carrefours vont lui permettre également une certaine transversalité d'exploration, à travers les documents.

L'indexation manuelle tient compte de ce critère d'accessibilité progressive à l'information en introduisant systématiquement des microtermes parfois assimilables à des divisions d'un plan de classement. Elle permet de situer le document dans un contexte plus général par rapport à des applications, à des domaines...

L'indexation Lexinet, quant à elle, offre une description plus spécifique du contenu du document, sans introduire systématiquement ce niveau supérieur de description que donnent les macrotermes. Pour les systèmes d'interrogation de bases de données, l'indexation Lexinet pourra être performante dans le cas d'une requête très spécifique, mais elle ne permet pas facilement un cheminement progressif dans l'ensemble des informations. . .

## 6.3. Cohérence d'indexation

La cohérence de l'indexation repose essentiellement sur la régularité de la pratique. Les documents étudiés ne peuvent être comparables que, si la description utilisée pour représenter leur contenu présente une variabilité minimum. Dans une perspective d'interrogation, la régularité d'indexation est nécessaire pour se garantir contre les silences éventuels qu'engendrerait au contraire une variabilité de description des documents. Or, par des techniques exclusivement manuelles, il est difficile de se prémunir contre cette variabilité car deux indexeurs ne retiennent pas toujours les mêmes termes d'un lexique pour décrire un même

document. Les méthodes automatisées ne présentent pas cet inconvénient, si ce n'est l'évolution de la langue et du vocabulaire de spécialité. Mais cette évolution n'est repérable que sur un laps de temps plus long.

## 6.4. Evolutivité

Pour l'interrogation des bases, l'évolutivité de l'indexation permet d'actualiser la description des documents en fonction des nouveaux problèmes qui sont posés. De façon plus générale, cette évolutivité permet de prendre en compte rapidement et rétrospectivement des évolutions des contenus scientifiques et techniques. Le concept d'évolutivité de l'indexation recouvre à notre avis deux notions :

- celle d'évolution des vocabulaires d'indexation,
- et celle de réindexation de chaque document (introduction, suppression, remplacement de termes).

L'évolution du vocabulaire est facilitée par des systèmes tels que Lexinet, qui s'appuient non pas sur un vocabulaire pré-établi, mais sur le vocabulaire extrait directement des textes [3].

La réindexation ne pose pas de problème dans les systèmes documentaires classiques quand il s'agit de supprimer ou remplacer un terme.

Des programmes informatiques peuvent être écrits pour modifier les enregistrements. Un problème plus délicat est la réindexation de documents anciens par un nouveau terme qui, auparavant, n'était pas admis dans le lexique. Avec les systèmes reposant sur une indexation manuelle, on ne peut réindexer facilement que les documents pour lesquels ce terme nouveau figurait en candidat descripteur. Des systèmes informatisés tels que Lexinet offrent la possibilité de faciliter ce travail: on pourrait réintroduire automatiquement un terme dans l'indexation des documents, à condition de garder dans un fichier toutes les adresses des termes (termes de l'antilexique). Cette possibilité pose tout de même un problème de stockage, car la quantité d'information à conserver risque alors d'être volumineuse.

## 7. Faut-il associer les deux modes d'indexation ?

Initialement, le système Lexinet a été conçu pour créer et gérer des listes de termes représentatifs d'un corpus de textes. Associé à des analyses statistiques, il nous a permis de rendre compte globalement d'un corpus de documents ; de nombreuses études ont déjà été menées dans cette perspective [11].

L'utilisation de Lexinet dans une perspective d'aide à l'indexation de documents en vue de servir un système rétrospectif impliquerait de mieux gérer les problèmes d'accessibilité et d'ambiguïté de l'information, deux points faibles de l'indexation Lexinet que nous avons mis en évidence précédemment.

Lever les ambiguïtés nécessiterait une étude de l'environnement des termes ambigus. Pour obtenir un niveau supérieur d'accessibilité, il faudrait introduire automatiquement des termes généraux soit par l'intermédiaire de relations sémantiques préalablement stockées dans l'ordinateur, soit en utilisant des méthodes statistiques de regroupement des termes (tels que les clusters construits par le logiciel LEXIMAPPE [3]). On permettrait ainsi un accès progressif à l'information tout en gardant un degré élevé de spécificité dans la description de chaque document.

Une autre façon de pallier les faiblesses de l'indexation Lexinet serait de l'associer à une supervision d'un expert humain. Cela reviendrait à associer les deux modes d'indexation, cherchant ainsi à unir les caractéristiques positives de chacune des indexations: les techniques Lexinet apporteraient un « plus » pour la régularité, la spécificité et l'évolution des termes, l'indexation manuelle permettrait d'introduire des descriptions intermédiaires, apportant ainsi une meilleure accessibilité à l'information, et levant les ambiguïtés possibles de certains termes.

*Mars 1989*

## Bibliographie

- [1] AUSTIN (D.) : Vocabulary control and information technology. – Aslib Proceedings, vol.38, n°1, 1986, p.1-15
- [2] CHARTRON (G.) : Analyse des corpus de données textuelles, sondage de flux d'informations. – Thèse de nouveau doctorat en traitement de l'information, Université de Paris-VII, juin 1988.
- [3] CHARTRON (G.) : Lexicon management tools for large textual databases : the LEXINET system. – A paraître.
- [4] FLUHR (C.) : Algorithme à apprentissage et traitement automatique des langues. – Thèse de doctorat ès sciences, Université Paris VI, juin 1977.
- [5] GASTALDY (B.) : De quelques éléments à considérer avant de choisir un niveau d'analyse ou un langage documentaire. – Documentation et bibliothèques, vol.32, n°1-2, 1986, p.3-23.
- [6] MICHELET (B.) : L'analyse des associations. – Thèse de nouveau doctorat en traitement de l'information, Université de Paris VII, octobre 1988
- [7] SALEM (A.) : Pratique des segments répétés, essai de statistique textuelle. – Publication de l'INALF, collection Saint-Cloud, Paris, 1987.
- [8] SALTON (G.) : The smart project. – Londres : Prentice Hall, 1971.
- [9] SALTON (G.), MCGILL (M.J.) : Introduction to moderne information retrieval. – International Student Edition, 1983.
- [10] SCHWARTZ (C.) : The TINA project : text content analysis at the central research laboratories at SIEMENS. – RIAO'88, Boston (USA), 21-24 mars 1988, p.361-368.
- [11] TURNER (W.A.), CHARTRON (G.), LAVILLE (F.), MICHELET (B.) : Packaging information for PEER review : new cword-analysis technic handbook of quantitation studies of science & technology, A.F.J. – Van Raan, Elsevier Science Publisher (North-Holland), 1988.
- [12] VAN RISJBERGEN (C.J.) : Information retrieval. – Butherworths, London, 1979.