LES EXPERIENCES D'INDEXATION A L'INIST

ROYAUTE J.
SCHMITT L.
OLIVETAN E.

Département Recherche et Produits Nouveaux
INIST - CNRS
2, Allée du Parc de Brabois
54514 VANDOEUVRE LES NANCY
FRANCE

ABSTRACT

We talk, in this paper, about the operation of indexation at INIST. We present two experiments carried out within the Department of Research and New Products that aim at the automation of indexation process. The first one comes within the scope of scientometric studies on text database. We have developped a software toolbox with which we can build a chain of treatments up to the generation of hyperdocuments. Therefore, indexation from a large corpus of source documents is the first module of that chain. In this part, we use linguistic and statistical methods to produce keywords from a stream of data. Linguitic heuristics are used to extract compound nouns or noun phrases from the text and combinational treatments determine the importance of each term according to the document. Keywords are here the input of an hypertext system. The second one is the development of a workstation for the information specialist integrating a computer-aided indexing system on title and abstract in bibliographical records. This indexing process works on a single bibliographical record and combines both linguistic methods and artificial intelligence (keywords generation). We use the same extraction module based on linguistic and add a knowledge based system to deduce implicit keywords. Finally, we show that the original specifications and purpose of each experiment are different and we start a discussion on the interest of these methods in relation to the kind of indexation wanted and the qualities expected from automatic indexing systems.

1 - L'INIST : producteur et consommateur d'indexation

L'INIST est une centrale documentaire qui produit deux bases de données (PASCAL et FRANCIS) couvrant l'ensemble des sciences et des techniques. Dans le but d'assister les Ingénieurs Documentalistes qui fabriquent ces deux bases, nous avons réalisé un prototype de station de travail ergonomique et convivial dont la principale fonctionnalité est une indexation assisté.

Les mots-clés de l'indexation peuvent être aussi le point d'entrée de systèmes informatiques fabriquant de l'information élaborée. C'est ainsi que nous construisons des hyperdocuments à partir de données bibliographiques ou textuelles pour des études bibliométriques et scientométriques [5]. Il est possible ainsi de regrouper les documents initiaux en classes sémantiques (soit les documents se référant à un même domaine, soit ceux issus d'un même laboratoire ou ville, soit ceux co-signés par les mêmes auteurs, etc ...) [6].

2 - L'indexation dans le dispositif d'analyse infométrique

Nous avons développé un ensemble d'outils permettant de passer d'un ensemble de textes non indexés à une structure hypertexte via un processus d'indexation automatique des textes et des mécanismes de clusterisation opérant un regroupement des mots-clés en classes (méthode des mots associés utilisant l'algorithme du simple lien [16]).

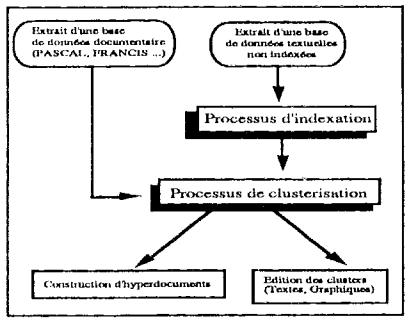

Fig. 1 : analyse infométrique

Le mécanisme d'indexation automatique que nous présentons dans cette section opère sur de gros volumes de texte et associe éventuellement des méthodes linguistiques à des traitements statistiques.

2.1 Méthodes statistiques d'indexation.

Ces outils d'indexation développés au département ont été élaborés à partir des méthodes statistiques présentées dans [1]. Ils reposent pour l'essentiel sur le repérage de mots ou de groupes de mots dans le corpus après suppression de "mots outils" : articles, prépositions, pronoms, etc ... Les traitements statistiques consistent en un filtrage distributif, par calcul de la variance de chaque objet et en un calcul d'une "fonction d'indexation" mesurant la représentativité mutuelle objet/document. Un seuil permet de retenir ou non cet objet comme pertinent du point de vue de l'indexation.

Ces méthodes sont intéressantes parce que simples à mettre en œuvre du point de vue de la programmation. Elles ne requièrent pas nécessairement l'accès à un dictionnaire électronique répertoriant la ou les appartenances syntaxiques de chaque mot. Elles présentent cependant l'inconvénient d'avoir des sorties bruitées dans la mesure où toute séquence de mots isolés pour l'indexation ne correspond pas nécessairement à un groupe nominal, mais peut inclure une séquence verbale plus ou moins figée (*N correspondant à N*, *N soumis à N*, *N semble* ...)[1] ou encore une locution adverbiale ou prépositionnelle figée (*de façon à, de manière à, en raison de*, etc ...). Ces deux types de séquences sont très nombreux dans les langues de spécialité. Cet inconvénient peut être limité, mais pas de façon significative, en ajoutant un anti-lexique de mots vides, c'est-à-dire ayant une sémantique faible du point de vue de l'indexation, parmi lesquels peuvent figurer : *façon*, *manière*, *raison*, etc ... Il faut cependant gérer cet anti-lexique : plusieurs passages du processus d'indexation sont nécessaires pour l'incrémenter.

Le principal problème dans les systèmes n'utilisant que des méthodes statistiques et combinatoires n'est donc pas la façon d'affecter des poids aux objets (termes simples ou pluritermes,) en vue d'éventuels filtrages sur leur distribution dans l'ensemble du corpus, mais la manière de reconnaître ces objets. En effet, les systèmes d'indexation ne faisant intervenir la linguistique qu'au niveau de la lemmatisation sur le genre et le nombre, segmentent le texte de façon souvent brutale : l'objet minimal étant le mot. La construction des pluritermes se fait par agrégation d'unitermes respectant des règles de proximité. Cette méthode de reconnaissance et de construction des objets de départ (sur lesquels vont s'effectuer les comptages) génère un bruit syntaxique important et accorde une place exagérée aux unitermes par rapport aux termes composés [16].

2.2 Méthodes linguistiques.

Afin de réduire les défauts inhérents aux méthodes statistiques et de récupérer un maximum de termes représentatifs de la terminologie du domaine, nous avons développé des procédures automatiques d'extraction reposant sur des observations linguistiques. Beaucoup de travaux sur l'indexation automatique sont fondés sur une extraction des groupes nominaux (*GN*). Ce choix est

[1] Dans la suite de cette exposé *N* désigne un nom, *Dét* un déterminant, *Prép* une préposition, *Adj* un adjectif, *Qua* un quantifieur, *V* un verbe conjugué. Le signe + indique une alternative, le symbole *E* correspond au mot vide.

justifié dans la mesure où pour les domaines de spécialité, l'information pertinente est localisée préférentiellement dans le *GN* [9]. Nous nous intéressons à ce sous-ensemble particulier des *GN* que sont les noms composés. Ils ont suscité un nombre important d'études linguistiques ces dernières années et on estime leur nombre, pour le français, largement supérieur à celui des noms simples.

Si les noms composés sont numériquement importants dans la langue ordinaire, c'est dans les langues de spécialité que leur fréquence est la plus élevée. Un dictionnaire terminologique ou un lexique d'indexation peut être considéré pour sa plus grande part comme étant un sous-ensemble des noms composés [17]. Le recours à la sémantique pour les identifier semble être une voie difficile. Le plus souvent leurs sens ne dérivent pas de la composition des parties du fait du figement : un *acier doux* fait référence à une variété particulière d'*acier* et l'adjectif *doux* ne spécifie en rien le nom *acier*. Des formes de surface identiques ont des sémantiques différentes : *N à N =: machine à vapeur* (1)
=: roue à aubes[2] (2)

Notre stratégie d'analyse s'appuie donc sur des observations linguistiques les concernant [7][8]. Ces différent travaux ainsi que nos propres observations montrent que les formes les plus productives dans la terminologie scientifique et technique sont du type :

N de N =: vitesse de corrosion
N Prép N =: corrosion sous tension
N Adj =: film passif

Contrairement à la stratégie de recherche des *GN* qui dépend essentiellement de la capacité à isoler différents syntagmes avec des grammaires adaptées, il n'existe pas de solutions algorithmiques permettant d'isoler les noms composés. Aussi sommes nous conduits à utiliser des heuristiques linguistiques, dont l'efficacité dépend étroitement du degré de figement des textes soumis à l'analyse. Ces heuristiques sont de trois types :

- Celles qui se fondent sur des séquences de la langue qui ne sont possibles qu'en présence de noms composés. De telles séquences peuvent être : *Dét Prép, Dét Qua, Dét N Qua, Dét V* comme dans *un à coup, un chasse goupille, un (moteur + E) deux temps*
- Celles qui identifient des mots séparés par un trait d'union : *ultra-vide, traction-tension, nickel-étain*, etc ...
- Celles qui s'appuient sur la probabilité de certaines suites à pouvoir être des noms composés. L'observation empirique montre que l'absence de déterminant après la préposition dans des suites *N Prép N* est un bon indice de figement (Ex: *corrosion par fatigue*). La juxtaposition de deux noms *N N* l'est également (Ex : *solution tampon*). Dans notre corpus de la métallurgie nous observons également un nombre important de formes de type *N à Dét N* qui sont également figées (Ex : *résistance à la corrosion*).

[2] Dans (1) il s'agit d'une *machine* qui fonctionne à la *vapeur*, dans (2) une *roue* constituée d'*aubes*.

Les principales phases du processus d'indexation sont les suivantes :

- Etiquetage syntaxique de tous les mots du texte par comparaison avec le dictionnaire électronique DELAF[3] du Laboratoire d'Automatique Documentaire et Linguistique (LADL) [3].
- Repérage de locutions adverbiales, prépositives ou conjonctives : *de manière à, de façon à, de sorte que ...*
- Réduction des appartenances syntaxiques multiples des mots, en alimentant un fichier de mots réduits à leur appartenance syntaxique la plus probable et aussi par l'utilisation de règles locales de désambiguisation.
- indexation par une recherche systématique, à l'aide de grammaires régulières, de formes syntaxiques indicatrices du figement des noms composés et les plus représentatives de la terminologie scientifique et technique. Les principales sont :

N1 # Prép # N2 # Prép # N3 # =: spectrométrie photoélectronique à rayons X
N1 Adj =: acier doux
N1 N2 =: solution tampon
N1 à Dét N2 =: stabilité à la corrosion

Le chiffre qui suit le symbole *N* indique sa position dans le groupe nominal. *N1* est toujours un nom. *N2* est un nom ou par défaut un mot non reconnu dans le dictionnaire. Le symbole # désigne une insertion optionnelle d'un adjectif, d'un nom ou d'un mot non reconnu ; signalons enfin que *Prép # N3 #* est optionnel. Nous proposons également comme candidat libre tous les noms simples qui ne s'insèrent pas dans ces schémas distibutionnels. La possibilité de traiter les mots non reconnus dans le dictionnaire et d'accepter des formes qui peuvent être ambigües confère au système une certaine souplesse.

Les résultats obtenus montrent que le bruit d'ordre syntaxique, dû essentiellement aux homographes ou aux mots non reconnus, est faible. Les termes de fréquence 1 représentent 85% de l'ensemble des termes extraits. Parmi eux, on peut distinguer deux groupes :

- des expressions figées, fortement représentatives du domaine : *spectroscopie de résonnance gamma.*
- des expressions non figées, mais souvent représentatives de la terminologie, dont le nom de tête a une occurrence forte : *solution de nitrate, solution de borate, solution fluorée*, etc ... L'indexation statistique ne retiendrait que le terme *solution*. Les termes de fréquence supérieure à un sont le plus souvent figés et pour la plupart représentatifs du domaine : *microscopie électronique à balayage, fissuration par corrosion sous tension*, etc ...

Nous mettons en évidence certaines régularités utiles à la recherche ultérieure d'informations :

[3] Le DELAF identifie 580 000 formes fléchies du français et leurs différentes appartenances syntaxiques (soit environ 80 000 formes lemmatisées). Pour l'anglais nous utilisons un autre dictionnaire électronique. Pour cette langue notre stratégie d'analyse s'apparente à celle de [12].

- Un petit nombre de noms simples apparaît très souvent dans les noms composés que nous isolons. Parmi les plus fréquents citons : *corrosion*, *alliage*, *acier*, *résistance*, etc ... Il est possible d'organiser une recherche intelligente autour de ces noms à partir des suites : *N N*, *N Adj*, *N Prép N* , etc ...

ex: *acier Adj* , permet de récupérer *acier doux*, *acier austénitique*, *acier inoxydable*.

- Les prépositions utilisées sont en nombre limité. Celles qui reviennent le plus souvent sont : *de*, *à*, *en*, *par*, *sous*. Il est possible de rechercher sélectivement les suites : *N de N* (*densité de courant*), *N à N* (*corrosion à chaud*), *N par N* (*zingage par immersion*), *N sous N* (*refusion sous laser*).

2.3 Applications.

Dans nos applications, nous cherchons à combiner les deux méthodes en fonction du volume de données à traiter. Les méthodes linguistiques que nous proposons peuvent être utilisées telles quelles sur de petits volumes ou dans le cadre d'une indexation "au fil de l'eau". Pour des volumes plus importants, les comptages deviennent pertinents. Il est donc intéressant d'utiliser les sorties du traitement linguistique comme entrées d'un traitement statistique. Le gain est double : premièrement le comptage se fait sur des éléments homogènes (des *GN* qui sont pour la plupart des noms composés), résultats du filtrage linguistique; deuxièmement le traitement statistique se limite ici à un calcul de la pertinence du mot-clé considéré en fonction de sa répartition dans le corpus et dans le document.

3 - L'indexation dans le cadre de la production d'une notice bibliographique

La principale tâche de l'ingénieur documentaliste (ID) dans la constitution des notices bibliographiques est la fonction d'indexation. Le processus d'indexation mis en œuvre dans ce projet utilise des méthodes linguistiques et des procédés d'intelligence artificielle de type "système à base de connaissances" [4].

3.1 Modélisation du processus d'indexation

Ce processus consiste, pour l'ID, à transcrire le contenu d'un document dans un langage documentaire après avoir extrait par analyse [14] les éléments d'information les plus significatifs pour une recherche ultérieure. Certaines parties du document, telles que le titre et le résumé, apparaissant comme les plus porteurs d'informations, sont privilégiées. Les termes apparaissant explicitement dans le document sont reconnus directement lors de la lecture. La lecture peut également "faire penser" à des mots-clés non explicitement décrits, qui font référence à la connaissance implicite du domaine qu'a le lecteur. Une fois cet ensemble de mots-clés (explicites et implicites) isolé dans le document, l'ID doit effectuer une sélection pour ne garder que ceux qui lui paraissent les plus pertinents dans le système documentaire, c'est-à-dire prioritairement ceux du lexique de référence PASCAL. Les termes appartenant à ce lexique sont appelés termes contrôlés. Ils peuvent être considérés comme les éléments de base de la connaissance de l'ID.

Le système que nous présentons ci-après peut être considéré comme une modélisation de la démarche logique d'indexation pour un domaine particulier. Cette modélisation s'est faite en collaboration avec les ingénieurs assurant la couverture du domaine des sciences de l'information (SI). Elle nous a conduits à structurer le sous-lexique des SI en une base de connaissances.

3.2 Organisation du système d'aide à l'indexation.

Nous avons réalisé un système d'aide à l'indexation interactif. Ainsi, à partir des termes contrôlés du titre et du résumé d'un document, nous générons un ensemble de termes pertinents pour l'indexation du document, parmi lesquels l'ID fait son choix pour réaliser l'indexation finale.

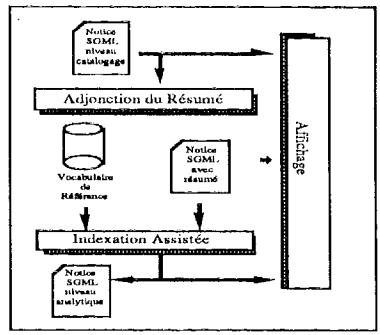

Fig. 2 : station de travail de l'ID

La reconnaissance de termes contrôlés apparaîssant dans le titre et le résumé du document à indexer s'appuie sur un pré-traitement linguistique. Les phrases du texte sont découpées en groupes nominaux qui sont lemmatisés puis comparés au lexique de référence PASCAL. Ce découpage en GN se fait à partir de mots ou séquences de mots se comportant comme des séparateurs dans la phrase. Il s'agit de séquences verbales, conjonctions, locutions (prépositionnelles, adverbiales ou conjonctives). Par comparaison avec le lexique de référence, nous obtenons ainsi une liste de termes contrôlés, appelés termes contrôlés explicites. Nous proposons aussi comme termes émergeants les termes extraits quand ils n'appartiennent pas au lexique PASCAL après application du module d'extraction des noms composés (cf. 2.2). L'ID a la possibilité d'insérer un ou plusieurs de ces termes s'il les jugent pertinents. Il a la possibilité de se créer et d'alimenter son propre lexique terminologique à partir de ce module d'extraction.

D'autre part, nous avons construit un thésaurus à partir d'un sous-ensemble du lexique PASCAL concernant les SI. La connaissance par l'ID de son domaine est ainsi

représentée par un ensemble de termes contrôlés et de concepts entre lesquels il existe des liaisons de différents types, stockés dans une base de faits de type thésaurus. La base de fait est organisée, dans son état initial sous forme d'arbre. Nous nommons concepts tous les noeuds de cet arbre, les feuilles sont, elles, des termes contrôlés. Chacun de ces concepts peut regrouper d'autres concepts (à des niveaux de profondeur quelconque) ou des termes contrôlés. Les relations dans cet arbre sont du type "générique-spécifique". La racine représente le concept général décrivant le domaine dans son entier (Fig .4). Un concept est décrit par un sous-ensemble du vocabulaire formé de termes contrôlés dont la signification est proche pour le domaine concerné. Nous pouvons ainsi proposer une catégorisation du vocabulaire de chaque domaine (les SI pour notre expérimentation). Les concepts de premier niveau décrivent les différents aspects les plus généraux du domaine traité et sont nommés index. Parmi ceux-ci un index particulier regroupe l'ensemble des thèmes principaux du domaine. Chaque index est lui-même redécoupé en sous-concepts ; le niveau de profondeur d'un concept dans l'arbre correspond au niveau de spécificité du concept pour l'index concerné. Il peut exister également des liaisons entre des concepts n'appartenant pas au même index : ce sont des liaisons d'ordre associatif dont la signification est "l'objet de départ fait penser à l'objet d'arrivée" [2]. Ces liaisons ne sont donc pas typées sémantiquement : elles ne dépendent pas d'un domaine précis. Elles serviront de support à la phase de déduction des concepts implicites. Ce type de liaison possède deux attributs : la force de la liaison et le niveau de propagation. Il peut se limiter aux seuls termes attachés au concept atteint ou concerner l'ensemble des termes attachés à tous les concepts fils du concept atteint.

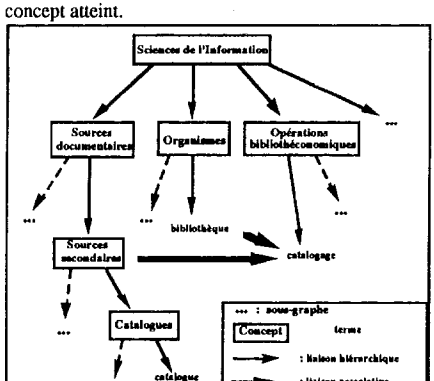

Fig. 3 : une partie de la base de faits

La structure finale retenue pour modéliser la base de faits est un graphe (Fig 3). Un nœud de ce graphe peut être de deux types : un terme contrôlé ou un concept. Un arc peut représenter deux liaisons différentes : une liaison hiérarchique ou une liaison associative pondérée.
Lors de la phase de déduction, les termes contrôlés extraits du titre et du résumé sont les entrées dans le graphe représentant la base de faits. Le mécanisme de déduction va activer l'ensemble des liaisons associatives dont l'objet de départ est un terme explicite, celles partant du concept auquel le terme appartient ainsi que les liaisons partant de ses concepts génériques. L'activation de cet ensemble de liaisons ne s'effectue qu'une seule fois à partir de la liste des termes explicites. Le mécanisme n'est pas réactivé sur les termes déduits.

Sur la figure 3, si les termes extraits sont *catalogue* et *bibliothèque*, alors le terme *catalogage* sera déduit directement par *bibliothèque* et par le concept *"Sources secondaires"*, concept ancêtre de *catalogue*. Sur cet exemple simple, on ne se préoccupe pas des pondérations des liaisons. Le résultat de cette phase est une prise en compte de l'ensemble de l'information apportée par les termes explicites en positionnant pour chaque terme déduit un certain nombre d'attibuts dont : le nombre de fois où le terme a été déduit quelque soit l'objet de départ, le nombre de fois où le terme a été déduit par une liaison partant d'un objet appartenant à l'index des thèmes, le nombre de fois où le terme a été déduit par une liaison forte.

Pour ne pas présenter l'ensemble des termes déduits de façon uniforme, on effectue un tri sur plusieurs des attributs de ces termes. Des heuristiques sur les valeurs des attributs, positionnés lors de la phase de déduction, permettent de proposer à l'intérieur d'un index deux listes de termes (éventuellement vides) : une liste principale où sont répertoriés les termes les plus pertinents et une liste secondaire qui contient des termes pouvant apporter des précisions intéressantes à l'ID.

Le domaine choisi pour tester le système d'aide à l'indexation est celui des Sciences de l'Information. Pour 2/3 des 200 documents testés, les termes de l'indexation manuelle se retrouvent entièrement dans les listes proposées par le système, 2/3 de ceux-ci étant répertoriés dans la liste principale. Le comportement du système est cohérent vis-à-vis des habitudes des ID du domaine et le nombre total de termes générés les satisfait pleinement.

4 - Méthode d'indexation et types d'applications

Notre propos n'est pas d'évaluer les différentes méthodes, mais plutôt d'étudier, à partir de leurs caractéristiques propres, comment les combiner en fonction des objectifs à atteindre. Schématiquement, nous pouvons distinguer deux stratégies d'indexation : celle réalisée au fil de l'eau (notice par notice) par les ID et celle portant sur un volume important de notices bibliographiques.

Nous distinguons dans le processus d'indexation au fil de l'eau, une phase d'extraction et une phase de génération. L'extraction concerne principalement l'identification de termes explicites et met en œuvre des traitements linguistiques. Elle consiste en une recherche des termes

du texte appartenant au lexique PASCAL et en une recherche des noms composés. Cette dernière est le garant de la spécificité de l'indexation et doit en outre permettre d'enrichir de façon incrémentale des dictionnaires terminologiques.

Alors que les heuristiques linguistiques de recherche de noms composés positionnent l'indexation du point de vue de la spécificité et de l'émergence de termes nouveaux pouvant être les futures vedettes d'un domaine scientifique, le couple extraction/génération de termes contrôlés se situe surtout dans la logique de son homogénéité. Les objectifs sont d'alléger la charge de travail de l'ID lors d'une analyse de document et de garantir une certaine cohérence à l'intérieur des bases bibliographiques, en guidant les indexeurs vers des choix similaires. Une certaine exhaustivité de l'indexation est également garantie en proposant des termes pour tous les aspects principaux du domaine traité.

Pour l'indexation de volumes importants de documents, nous pouvons choisir ou combiner des méthodes statistiques ou linguistiques. Ces traitements sont souvent la première étape d'un traitement statistique permettant de produire une information élaborée par des méthodes de clusterisation ou de classification (§ 1, § 2). Ces méthodes peuvent donner lieu à des produits de type hypertextes permettant de naviguer dans un réseau de mots-clés et d'accéder aux références bibliographiques. Elles peuvent être le support d'études scientométriques ou un outil d'évaluation de l'indexation produite. L'intérêt peut se porter aussi sur les termes de fréquence 1, que seules les méthodes linguistiques peuvent faire émerger. Une méthodologie reste à mettre en place, qui différencie des formes non figées et très productives (*solution de borate, solution de HCl, ...*) pouvant être regroupées sous un vocable plus générique, des formes figées (*solution tampon, ...*).

5 - Conclusion

Cette présentation des différentes aides à l'indexation que nous proposons tend à montrer qu'il n'y a pas une stratégie unique pour résoudre le problème, mais différentes approches adaptées à des besoins particuliers, celles-ci apparaissant souvent complémentaires. C'est la raison pour laquelle nous nous efforçons de ne pas développer des produits monolithiques, difficiles à maintenir, mais des modules outils pouvant se combiner en fonction du type d'application concerné.

[1] CHARTRON G.
Analyse des corpus de données textuelles, sondage de flux d'informations - Thèse de nouveau doctorat en traitement de l'information - PARIS-VII, Juin 1988.

[2] CHAUMIER J.
Le traitement linguistique de l'information - Entreprise Moderne d'Edition, PARIS (F), pp. 149-162, 1988.

[3] COURTOIS B.
Un système de dictionnaires électroniques pour les mots simples du français - Langue française N° 87, PARIS (F), 1990.

[4] DISCROLL P.H.
The operation and performance of an artificially intelligent keywording system - Information Processing & Management Vol. 27, N° 1.

[5] DUCLOY J. ; GRIVEL L. ; LAMIREL J.C. ; POLANCO X. ; SCHMITT L.
INIST'S experience in hyper-document building from bibliographic databases - Conférence RIAO 91 BARCELONE (SP), 2-5 avril 1991.

[6] GRIVEL L. ; LAMIREL J.C.
SDOC, a generator of hypertext structures - 2nd Conference Multimedia Information - CAMBRIDGE (UK), 15-18 juillet 1991.

[7] GROSS G.
Structure des noms composés - Colloque Informatique et Langue naturelle - NANTES (F), 12-13 octobre 1988.

[8] GROSS G. ; DUGAS A.
Analyse des groupes N de N - Colloque Informatique et Langue naturelle - NANTES (F), 23-24 janvier 1991.

[9] GROSS M.
Les industries de la langue et l'étude du français - Langue Française n°83, pp. 88-100, PARIS (F), 1989.

[10] GROSS M.
La construction de dictionnaires électroniques - Ann. Télécommun., tome 44, n° 1-2, pp. 4-19, 1989.

[11] GROSS M.
Les Banques de données du LADL. Analyse automatique du français et couverture - Colloque Informatique et Langue naturelle - NANTES (F), 23-24 janvier 1991.

[12] KLINGBIEL P.H.
A technique for machine-aided indexing - Information Storage and Retrieval, vol 9, pp 477-494 - Pergamon Press, 1973.

[13] LAPORTE E.
Reconnaissance des expressions figées lors de l'analyse automatique. - Langages n° 90, Larousse, pp 117-126, PARIS (F), 1988.

[14] MENILLET D.
Règles d'indexation pour la base de données bibliogaphiques PASCAL - INIST PARIS, 1990.

[15] MICHELET B.
L'analyse des associations - Thèse de nouveau doctorat en traitement de l'information - Université de PARIS-VII, Octobre 1988.

[16] POLANCO X. ; SCHMITT L. ; BESAGNI D. ; GRIVEL L.
A la recherche de la diversité perdue : est-il possible de mettre en évidence les éléments hétérogènes d'un front de recherche ? - Journées d'étude sur les systèmes d'information élaborée : bibliométrie, information stratégique, veille technologique - ILE ROUSSE (F), 2-5 juin 1991.

[17] WAGNER H.
Dictionnaires, bases de données lexicales et lexicographie des langues de spécialité : le traitement des unités complexes - Colloque Informatique et Langue Naturelle - NANTES (F), 23-24 janvier 1991.