Caractérisation de parties de discours scientifiques : Analyse des corrélations entre propriétés

Christine Michel, Eric Guinet, Thierry Lafouge

Caractérisation de parties de discours scientifiques : Analyse des corrélations entre propriétés.

Christine Michel, Eric Guinet, Thierry Lafouge
michel@recodoc.univ-lyon1.fr, quinet@recodoc.univ-lyon1.fr, lafouge@enssibhp.enssib.fr

Laboratoire RECODOC,
Bat 721, Université Claude Bernard LYON I
43, Bd du 11 Novembre 1918
69622 VILLEURBANNE CEDEX
Tel : 04 72 43 13 91

Résumé

La régularité de structure des articles en information scientifique et technique nous a permis de définir des propriétés de caractérisation des parties de discours, nous les appelons unités documentaires. Nous avons choisi de les décrire selon leur *type* (introduction, résultats, conclusion …), leur *forme discursive* (argumentatif, narratif, …) ou leur *style de présentation* (littéraire, données numériques, schéma, …). Ces propriétés utilisées dans le cadre du projet Profil-Doc, permettent de construire une base de données où l'information extraite lors d'une interrogation est filtrée selon une fonction d'usage pour l'utilisateur. L'article suivant propose des résultats expérimentaux sur la caractérisation d'une telle base de données. Nous avons en particulier cherché à vérifier s'il existe (ou non) une régularité des propriétés des unités documentaires ainsi que d'éventuelles corrélations entre elles. Cette analyse s'appuie sur une étude statistique de type AFC (Analyse factorielle des correspondances).

Introduction

Pour pallier aux limites de l'"indexation" et avoir une meilleure connaissance du fonds, les systèmes documentaires traditionnels et automatiques décrivent les documents par des critères externes à leurs contenus : le pays, l'année de publication, le nom du laboratoire, etc. Le choix de propriétés de description est la source de nombreuses réflexions. Dans la plupart des cas, les propriétés sont utilisées pour décrire le document dans sa globalité. Une étude approfondie [BEN97] sur certain nombre de textes, livres, thèses, articles de revues scientifiques a montré qu'un texte possède une *structure générale*. Il forme une unité car il est construit pour faire passer un message : résultats d'expérience, synthèse, nouvelles pistes de recherche, etc. Cette unité matérielle et intellectuelle est le résultat d'un lien parfaitement établi entre ses différentes parties, celles-ci pouvant former à leur tour des unités indépendantes remplissant une fonction bien déterminée. Cette constatation nous a conduit à admettre que "l'éclatement" du document selon ces unités nous permet, tout en préservant l'unité globale du document, de présenter à l'utilisateur une information plus affinée et plus facile à saisir [LAI94].
Dans cette optique, le système Profil-Doc[1] [LAI96] se propose d'aller plus loin que la simple utilisation de ces critères pour la description et sélection des documents. Ces propriétés nous

[1] Profil-Doc est un système de recherche d'information documentaire en texte intégral développé par le laboratoire RECODOC.

permettront de sélectionner un corpus "personnalisé" suivant des caractéristiques de l'utilisateur, corpus sur lequel portera la question. En d'autres termes, ces propriétés, appariées avec le profil de l'utilisateur, nous permettent de présélectionner un ensemble de parties de documents, nous les appellerons unités documentaires. **La manière selon laquelle sont attribuées les propriétés diffère complètement du processus de l'indexation.**

1- La base de données d'unités documentaires

Nous avons construit une base de données en texte intégral dans laquelle les documents sont découpés en unités documentaire caractérisées selon ces propriétés[2]. Le découpage et la caractérisation des documents ont été fait manuellement. A l'heure actuelle la base comprend 505 unités documentaires extraites de 55 documents [MIC99]. Les articles proviennent de revues de presse professionnelle, de revues de presse fondamentale et d'ouvrages en Sciences de l'information et de la communication[3] , en biomécanique[4], en biologie[5] et en pharmacie[6]
Deux études ([BEN97] [BEN97B]) sur la fonction des différentes parties de discours, menées dans le cadre du laboratoire RECODOC ont visé à valider le choix des différentes propriétés de description des unités documentaires. Dans le tableau suivant nous pouvons voir les propriétés sélectionnées ainsi que la répartition de chacune d'elle dans notre base .

Propriétés propres au document			**Propriétés propres à l'unité documentaire**					
1- Type d'environnement éditorial			*5- Style*			*7- Type d'UD*		
Divers	11	2,2 %	Calculs	16	3,1%	Annexe	18	3,6 %
Mémoire 3ème cycle	32	6,3 %	Données numériques	8	1,6 %	Bibliographie	33	6,5 %
Presse fondamentale	171	33,9 %	Littéraire + calculs	11	2,2 %	Conclusion	40	7,9 %
Presse professionnelle	291	57,6 %	Littéraire + données numériques	63	12,5 %	Contexte	70	13,9 %
2- Profession de l'auteur						Développement	63	12,5 %
Etudiant	53	10,5 %	Littéraire	369	73,1 %	Discussion	58	11,5 %
Spécialiste	452	89,5 %	Représentation	38	7,5 %	Environnement	9	1,8 %
3- Communauté de l'auteur			*6- Forme discursive*			Expérimentation	25	4,9 %
Etudiant	6	1,2 %	Argumentatif	157	31,1 %	Introduction	44	8,7 %
Industriel	18	3,6 %	Descriptif	343	67,9 %	Méthode	45	8,9 %
P.M.E.	4	0,8 %	Discours rapporte	4	0,8 %	Résultats	63	12,5 %
Public ou parabublic	82	16,2 %	Narratif	1	0,2 %	Résumé	29	5,7 %
Universitaire	395	78,2 %				Thème	8	1,6 %
4- champ disciplinaire de l'auteur								
Biologie	100	19,8 %						
Physique	34	6,7 %						
SIC	371	73,5 %						

Tableau 1 : Répartition des différentes propriétés

Lorsque nous avons défini les propriétés propres à l'unité documentaire -: le type, le style et la forme de l'UD - nous avons fait l'hypothèse qu'elles étaient présentes de façon indépendante les unes des autres. Notre objectif est d'observer si cette hypothèse se réalise. Si c'est la cas, nous pourrons accorder une validité au traitement opéré pour découper et caractériser les

[2] Nous ne passons pas par une étape de compréhension du contenu pour attribuer les propriétés. En effet, elles sont soit facilement repérables à l'intérieur du document ; c'est le cas pour la propriété *style de l'unité documentaire*, soit repérées par certains marqueurs (linguistiques ou autres), cela peut être le cas pour la propriété *forme discursive* de l'unité documentaire.
[3] BBF, Documentaliste, La revue française de bibliométrie, Laforia, Cahiers du Lerass, IDT 96
[4] AUTOMEDICA, Archives de physiologie et de biochimie, Congrès de la société de biomécanique, ITBM, Conference of medical and Health Libraries
[5] Biologie Terminale D, C. R. Acad, articles de recherche disponibles sur Internet
[6] Journal de Pharmacie de Belgique

unités. Sachant que ces opérations sont manuelles, nous ne sommes pas certain que le jugement, subjectif, de la personne qui effectue l'opération n'introduise pas de biais.
Nous observerons les distributions des propriétés des unités documentaires ainsi que d'éventuelles **corrélations entre elles** par deux études statistiques : le test du c^2 et l'AFC.

2- Analyse bidimensionnelle : test de dépendance entre propriétés

Nous avons croisé deux à deux les propriétés suivantes, *Type d'unité documentaire, Forme discursive de l'unité documentaire, Langage (style) de l'unité documentaire*, pour observer s'il y avait ou non indépendance. Dans les trois tableaux suivants (Tableaux 2, 3 et 4) nous présenterons les résultats des c^2 totaux (dernière case en grisé) ainsi que les contributions relatives de chaque case (c^2 **par case**) au calcul du c^2. Les justifications théoriques de calculs sont présentées dans [SPI76]. L'interprétation des résultats se fait a l'aide de la table du c^2. Si la valeur calculée est supérieure à la valeur lue dans la table, avec le même degré de liberté et pour un seuil de signification déterminé, on peut conclure que les variables sont dépendantes. Si la valeur du c^2 établit qu'il y a un lien, les contributions relatives nous permettrons d'affiner l'analyse en mettant en évidence les couples de modalités qui contribuent plus particulièrement au calcul du c^2 [CIB83].
Dans le tableau 2 croisant le type de l'UD avec la forme discursive, nous observons une valeur de 295,02 qui dépasse largement la valeur lue dans la table du c^2 (au seuil de signification 0,005 la table donne $c^2_{0.995} = 66{,}8$ avec 36 degrés de liberté). De la même manière dans le tableau 3, croisant le type de l'UD avec son style, nous observons une valeur de 684,2, alors que sous les même hypothèses, avec 60 degrés de liberté, la table donne $c^2_{0.995} = 92$. Ces deux valeurs très fortes de c^2 total signifient que les propriétés sont dépendantes. Observons plus précisément comment contribuent chacun des couples de modalités de propriétés au c^2 total.

χ^2calculés	Argumen-tatif	Descriptif	Discours rapporté	Narratif	Total
Annexe	1,204	0,630	0,143	0,036	2,012
Bibliographie	10,259	5,000	0,261	0,065	15,586
Conclusion	**48,522**	**21,500**	0,317	0,079	**70,418**
Contexte	1,526	0,876	0,555	0,139	3,095
Développement	5,722	2,937	0,499	0,125	9,282
Discussion	**71,747**	**39,390**	**27,287**	0,115	**138,540**
Environnement	2,798	1,364	0,071	0,018	4,251
Expérimentation	5,901	2,902	0,198	0,050	9,051
Introduction	1,365	0,799	0,349	**9,564**	12,076
Méthode	5,777	2,913	0,356	0,089	9,136
Résultat	5,722	2,937	0,499	0,125	9,282
Résume	7,127	3,500	0,223	0,057	10,914
Thème	0,920	0,378	0,063	0,016	1,378
Total	**168,59**	**85,13**	30,83	10,48	**295,02**

Tableau 2 : Croisement des propriétés Forme discursive et Type de l'UD

χ^2 calculés	Calculs	Données numériques	Littéraire + calculs	Littéraire + données numériques	Littéraire	Représen-tation	Total
Annexe	3,584	0,285	0,392	10,067	1,311	1,355	14,328
Bibliographie	1,046	0,523	0,719	4,117	**24,113**	**375,040**	**405,557**
Conclusion	1,267	0,634	0,871	4,990	3,970	3,010	7,762
Contexte	2,218	1,109	0,148	6,847	4,913	5,267	10,322
Développement	0,505	0,998	0,287	0,166	0,191	4,741	1,956
Discussion	1,838	0,919	1,263	1,447	2,661	2,594	5,468
Environnement	0,285	0,143	0,196	0,013	0,027	0,154	0,637
Expérimentation	1,842	0,396	3,890	**15,182**	2,891	1,881	21,310
Introduction	1,394	0,697	0,958	5,489	4,367	3,311	8,539
Méthode	**30,315**	0,713	**9,303**	0,464	1,440	0,044	40,795
Résultat	0,497	**49,130**	0,101	**46,615**	8,719	4,741	**96,338**
Résume	0,919	0,459	0,632	3,618	2,878	2,182	5,628
Thème	0,254	0,127	0,174	0,998	0,794	0,602	1,552
Total	45,96	56,13	18,94	**100,01**	58,28	**404,92**	**684,23**

Tableau 3 : Croisement des propriétés Langage et Type de l'UD

χ^2 calculés	Calculs	Données numériques	Littéraire + calculs	Littéraire + données numériques	Littéraire	Représen-tation	Total
Argumentatif	0,7836	2,487	0,0984	3,7640	4,3275	8,1524	19,6131
Descriptif	0,4185	1,212	0,0297	1,9823	2,2275	4,0232	9,8933
Discours rapporté	0,1267	0,063	0,0871	0,4990	0,3970	0,3010	1,4743
Narratif	0,0317	0,016	0,0218	0,1248	0,0993	0,0752	0,3686
Total	1,3605	3,778	0,2371	6,3701	7,0513	12,552	**31,3492**

Tableau 4 : Croisement des propriétés Forme discursive et Langage de l'UD

Dans le tableau 2 nous pouvons observer que la plus grande contribution au c^2 est donnée par le couple [discussion, argumentatif] pour 71,747 soit 24,31%. Viennent ensuite des couples de contributions relatives au c^2 total moindre: le couple [conclusion, argumentatif] pour 16,44%, [discussion descriptif] pour 13,35%, et [discussion, discours rapporté] pour 9,24%.

Ces dépendances s'expliquent assez bien si l'on **considère la fonction informationnelle de chaque partie de discours**. Si nous reprenons l'étude faite dans [BEN97] sur la définition des différents types d'unités documentaires, nous voyons que pour remplir cette fonction certaines parties de discours sont écrits avec une forme discursive bien définie. Ainsi, les discussions et conclusions (qui sont souvent liées) sont écrites dans un style argumentatif pour, à partir de résultats, convaincre le lecteur de la réfutation ou de l'acceptation d'une nouvelle théorie ou du bien fondé d'explorations complémentaires. Rencontrer une dépendance des discussions avec les styles descriptifs et discours rapportés est lié à cette même nécessité de convaincre : les parties descriptives peuvent par exemple présenter des travaux antérieurs, en opposition ou en soutien de l'étude, pour la renforcer ; les discours rapportés peuvent être des citations.

Dans le tableau 3, la contribution maximale, soit près de 60%, vient du couple [bibliographie, représentation]. Ce couple est à mettre à part car, plus qu'une dépendance la bibliographie est par définition normalisée, et donc à classer systématiquement dans un langage de type représentation.
Les autres contributions représentatives au c^2 total proviennent des couples [résultats, données numériques] pour 7%, [résultats, littéraire + données numériques] pour 6,8%, [méthode, calculs] pour 4,4%, [expérimentation, littéraire + données numériques] pour 2,2% et [annexe, littéraire + données numériques] pour 1,4%.
La nature des informations présentées explique bien les dépendances observées. Les résultats sont souvent des données numériques présentées ou non avec un texte explicatif. Les annexes sont souvent des présentations de résultats et les expérimentations des grilles pour la lecture des résultats, ce qui explique la forte proportion d'unités de type «littéraire contenant des données numériques». Enfin la méthode est souvent une description formelle sous forme d'équations d'une théorie, ce qui explique la dépendance avec les unités de type calculs.

Dans le tableau 4 nous pouvons observer un résultat de 31,3. Selon la table du c^2, en prenant un risque de $5^0/_{00}$, nous avons une valeur de seuil $c^2_{0.995}$ = 32,8 avec 15 degrés de liberté. En prenant un risque de $5^0/_0$, nous avons une valeur de seuil $c^2_{0.95}$ = 25. Nous somme donc au dessus du seuil à $5^0/_0$. Nous pouvons cependant dire que si dépendance il y a, cette dépendance est beaucoup moins marquée que dans les deux cas précédents.
Les tableaux 2 et 3 nous montrent des dépendances relatives entre quelques modalités de propriétés. Les c^2 totaux forts permettent de supposer une dépendance entre ces propriétés. Une AFC va nous permettre de dégager d'une manière plus fine comment s'illustre cette dépendance.

3- Analyse multidimensionnelle : Visualisation cartographique des dépendances

L'AFC permet de réduire à un espace à deux ou trois dimensions, l'espace de représentation des éléments des ensembles analysés. Cette réduction sera opérée en fonction des proximités entre les diverses unités documentaires et les modalités de propriétés. La proximité sera calculée en utilisant la distance usuelle du c^2. Elle permettra d'avoir une représentation graphique des calculs précédents.
Pour procéder à l'analyse des résultats, nous interpréterons la proximité des points aux axes ainsi que celle des points entre eux.

Les éléments permettant la validation de l'interprétation sont les paramètres usuels mis en évidence par [BEN73]:
- L*es valeurs propres* : elles assurent la bonne représentation barycentrique des points le long de l'axe correspondant (plus elles sont proches de 1 et meilleure est la représentation),
- *Les contributions à l'inertie* (ou contributions absolues) : elles décrivent la part prise par un élément dans la construction d'un axe,
- *Les cosinus carrés* (ou contribution relative) : ils mesurent la qualité de la représentation de chaque élément par les axes.
L'ensemble des résultats numériques sont présentés dans une thèse [MIC99], nous ne présentons ici que les visualisations graphiques.

a) AFC du croisement des propriétés « Type d'UD » et « Forme discursive d'UD »

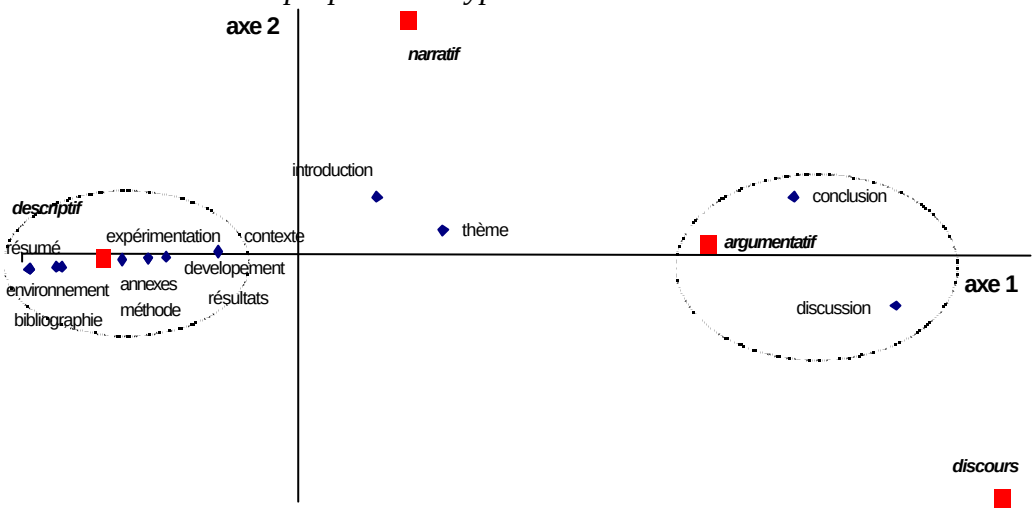

Figure 1 : Axes 1 et 2 de l'AFC pour les propriétés Type et Forme discursive d'UD

L'axe 1 explique 91,47% de l'information. Il montre l'opposition qu'il peut y avoir entre des unités documentaires argumentatives et descriptives. Sur la droite de l'axe 1 (Figure 1), on retrouve bien la dépendance entre les discussions et conclusions argumentatives remarquée avec les calculs du c^2. A l'opposé, sur le même axe, un groupe de points représentant le reste des unités documentaires, est centré autour de la propriété « descriptive ». Si la dépendance entre la forme discursive «argumentative» et les types « conclusion » et « discussion » est nettement visible, il n'est pas évident d'en dire autant du groupe de points centrés autours de la propriété «descriptive ». En effet, on peut considérer que par défaut, un texte qui n'est pas argumentatif est la plupart du temps descriptif.
L'axe 2 est beaucoup moins explicatif (5,56%). Il montre l'existence d'une opposition sur la nature tantôt narrative ou tantôt en discours rapporté des introductions ou des descriptions de thème (et dans une moindre mesure les conclusions et discussions). Ceci est tout à fait représentatif des deux moyens qu'a un auteur pour récapituler des travaux ou théories antérieurs. Il peut en effet citer directement les autres auteurs, ce qui correspond à un discours rapporté, ou bien énumérer l'évolution de leurs recherches, ce qu'il fait dans un style souvent narratif.
Cette AFC nous permet de voir émerger les trois fonctions informationnelles principales d'un discours scientifique. Nous voyons apparaître en premier la fonction de présentation, de description et d'explication d'un sujet. Dans des parties de textes comme l'introduction ou la description de thème, l'auteur pose les bases de son étude. La deuxième fonction d'un discours scientifique est de présenter un travail de recherche particulier, innovant. C'est ce

que l'auteur fait dans le corps de son texte. Sa troisième étape est d'en tirer certaines conclusions et réflexions, qu'il expose sous forme argumentative dans la discussion ou la conclusion. Il essaie ainsi de convaincre le lecteur de la validité de son travail personnel.

b) AFC du croisement des propriétés « Type d'UD » et « Style »

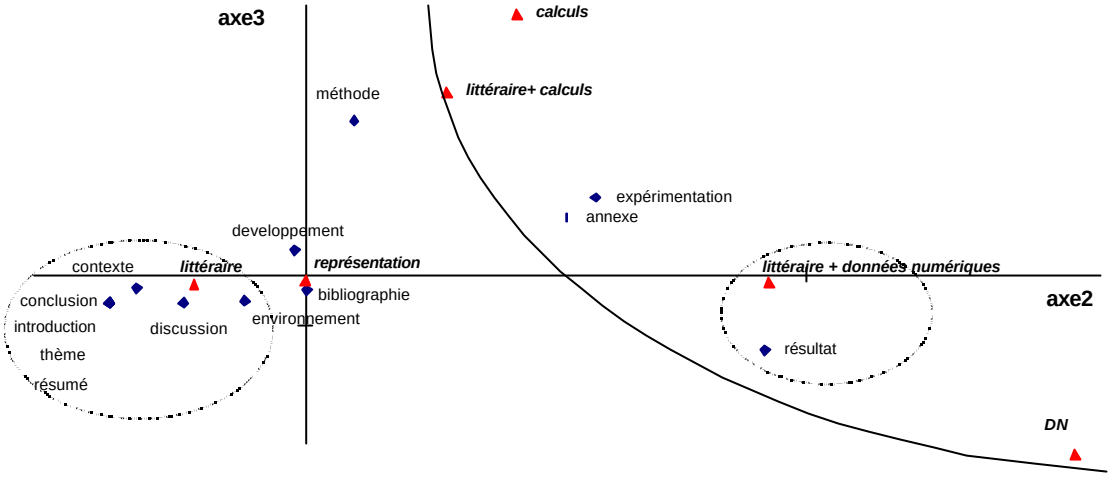

Figure 2 : Axes 2 et 3 de l'AFC pour les propriétés Type et Style d'UD

L'axe 1, non représenté sur la figure 2, explique 64% de l'information. Nous ne l'avons pas représenté car il n'est cependant représentatif que d'un lien ; celui qui existe entre les modalités « représentation » et « bibliographie ». Nous avons déjà noté et expliqué cette dépendance avec les calculs du c^2 (C.F. tableau 3). Nous présentons donc le schéma de l'AFC avec les axes 2 et 3 car ceux-ci, bien que moins explicatifs, apportent de l'information complémentaire.

L'axe 2 de la figure 2 explique 23,8% de l'information. On peut y voir l'opposition qui existe entre : les parties de textes qui sont purement littéraires sur la gauche du schéma, celles à droite contenant des données numériques, des descriptions intermédiaires où les données numériques sont présentées avec une analyse explicative sous forme de texte littéraire et des calculs. Les parties de texte composées exclusivement ou en partie de données numériques sont : les expérimentations, les annexes et les résultats. L'axe 3 présente l'opposition entre des parties de texte présentant des calculs et celles contenant des données numériques. On y retrouve principalement les parties « méthodes » présentant les calculs.

La grande majorité des UD sont de style «purement littéraire ». Ceci s'explique facilement de part la nature de l'information présentée, si l'on considère qu'une grande partie des articles composant la base ont été extraits de revues où l'information est explicité sous forme de texte (en opposition à des revues où l'information est présentée directement sous forme d'équations ou d'algorithmes).

Conclusion

L'étude ci-dessus montre qu'il existe effectivement des dépendances entre les propriétés Forme discursive et Type d'UD d'une part, et Style et Type d'UD d'autre part. Ceci semble aller à l'encontre de notre hypothèse d'indépendance. Toutefois nous pouvons précisez que les dépendances observées ne sont pas globales à toutes les modalités de chaque propriété mais seulement à quelque unes d'entre elles. Rappelons (Tableau 1) que la base est essentiellement constituée d'unités documentaires extraites de revues de presse fondamentale ou professionnelle, dont les auteurs sont souvent des universitaires spécialistes dans le

domaine des sciences de l'information. Ce type d'écrits scientifiques a été largement analysé dans [BENA97]. Or, les dépendances entre modalités correspondent précisément à des régularités de rédaction qui sont facilement explicables. Avec le calcul du c^2, nous avons vu que la dépendance entre certaines modalités des propriétés Forme discursive et Type d'UD était liée à **la fonction informationnelle** des parties de discours. La conclusion et la discussion par exemple, seront souvent écrites dans une forme discursive argumentative car leur rôle est précisément de convaincre le lecteur. L'AFC nous a de plus permis de voir émerger les trois fonctions informationnelles principales d'un discours scientifiques: la présentation générale du sujet, la description de l'étude réalisée, et l'argumentation de l'auteur. Cette dernière sert au lecteur dans son acceptation de la validité du travail de recherche. La dépendance entre certaines modalités des propriétés Style et Type d'UD est lié à **la nature** de l'information présentée. Il est par exemple évident que des résultats numériques seront présentés sous forme « données numériques » ou éventuellement « littéraires + données numériques » si l'auteur prend le soin d'y adjoindre des explications ou des analyses complémentaires.

Bibliographie

[BEN97] BEN ABDALLAH N – Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information utile : vers un système d'information évolué. – Thèse soutenue le 7 juillet 1997 à l'Université Lyon I.

[BEN97B] BEN ABDALLAH N, MICHEL C., LAINE-CRUZEL S. - LAFOUGE T. - *Caractérisation et découpage de textes scientifiques pour la construction de systèmes de requête personnalisés* – Actes du congrès Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF « L'ingénierie de la langue : de la recherche au produit ». Avignon, 15-16 avril 1997.

[BEN73] BENZECRI – L'analyse des Données tome II – L'analyse des correspondances. Edition Dunod – 1973 – 619 p.

[CIB83] CIBOIS P. - L'analyse factorielle. Edition Que sais -je? - 1983

[LAI94] LAINE-CRUZEL S. - *Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs.* Documentaliste. Sciences de l'information - 31 (3) – 1994 - pp 143-147.

[LAI96] LAINE-CRUZEL S., LAFOUGE T., LARDY J.P., BEN ABDALLAH N. - *Improving information retrieval by combining user profile and document segmentation.*- Information Processing and management -1996- vol 32 n 3 - pp 305-315.

[MIC99] MICHEL C - Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et évaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs.- Thèse soutenue le 6 janvier 1999 à l'Université Lyon II -322 p.

[SPI76] SPIEGEL M. R. - Théorie et applications de la statistique - Série Schaum Groupe McGraw-Hill - Ediscience – Paris - 1976 - 357 p.