La Segmentation des documents techniques en amont de l'indexation : définition d'un modèle

OUERFELLI Tarek
Institut Supérieur de Documentation
Campus Universitaire de la Manouba
2010 - BP 600
Tunisie

1- Introduction

Dans les systèmes documentaires classiques, l'utilisateur reçoit le plus souvent les références des documents et au mieux les documents primaires et c'est à lui de dépouiller le document pour juger s'il répond à son besoin ou non. Ainsi, la demande d'information sera transformée en demande de documents. Le document contient l'information, mais cette dernière est trouvée indirectement.

Ce genre de réponse, dans un processus de recherche d'information dans un document technique, n'est pas adapté à la situation. L'utilisateur du document technique, qui est souvent chargé d'exécuter les procédures et de maintenir le dispositif en état de marche, recherche de l'information en vue de répondre à un besoin professionnel. Il effectue une recherche dans le but de savoir pour faire, tout en cherchant généralement à atteindre directement l'information la plus élémentaire satisfaisant son besoin. Le processus de recherche d'information se doit alors d'être particulièrement rapide et efficace, d'où vient l'intérêt de traiter le document technique comme une construction moléculaire. Il sera décomposé pour donner naissance à de nouvelles unités utilisables et ce pour effectuer des lectures spécifiques. Ce qui permettra, d'une part, une représentation fine de son contenu, et d'autre part, un accès plus localisé à l'information facilitant la tâche de consultation pour l'utilisateur (Salton et al 96).

Par conséquent, l'indexation du document technique nécessite une étape préliminaire consistant à le segmenter en unités fines. C'est à cet aspect que nous allons nous intéresser dans cet article, en proposant un modèle qui part de la réalité des documents techniques.

2- Propriétés des documents techniques

Nous désignons par document technique, les documents du type manuel d'utilisation de dispositifs techniques complexes. Ce document véhicule des savoirs et des savoir-faire propres à un champ technique particulier. Il représente aussi bien la description d'une machine (avion, train, système informatique,...), du fonctionnement de cette machine et des divers processus la concernant, que la description des procédures de réalisation d'une action technique dans un environnement bien précis. L'objectif de l'utilisation du document technique est essentiellement à visée opératoire, pour réaliser une tâche ou une action (Vigner et al 76, Bronckart 85).

Le document technique, qui est généralement volumineux, se caractérise par une forte structuration avec une organisation logique bien définie. Du point de vue linguistique, il a des traits qui lui sont propres : la grammaire de la phrase est simple, le lexique est monosémique ; il désigne sans la moindre ambiguïté telle pièce, tel outil ou telle opération. Ce caractère univoque et monoréférentiel des termes du vocabulaire véhiculé par le document technique se

reconnaît au fait qu'il est impossible de substituer un terme à un autre. Ainsi, chaque objet serait représenté par un terme unique et chaque terme n'a qu'un sens dans le domaine concerné.

Les documents technique sont aussi hétérogènes. Cette hétérogénéité s'exprime par la pluralité des modes véhiculant l'information. On trouve, outre les séquences textuelles écrites dans la langue naturelle, des énoncés et des séquences de commandes écrits dans un langage formel (langage informatique). L'hétérogénéité des documents techniques s'exprime aussi par la présence d'objets non textuels, en l'occurrence les figures et les tableaux. Cette pluralité, comme l'ont fait remarquer C. Froissart et G. Lallich (99), ne se présente pas sous la forme d'alternance mais de cohabitation.

3- La segmentation pour l'indexation : *Etat des lieux*

La tâche de segmentation des textes est traitée sous plusieurs angles dans la littérature selon la finalité visée : reconnaissance du texte ou bien extraction et recherche d'information. Dans le domaine de la recherche d'information, on distingue différentes méthodes de segmentation :

- Segmentation en une suite de mots ;
- Segmentation en phrases ;
- Segmentation en paragraphes ;
- Segmentation thématique ;
- Segmentation en unités logiques répercutées dans le sommaire.

Ces différentes méthodes sont présentées en détail dans Ouerfelli (01). Nous allons nous contenter ici à présenter notre point de vue par rapport à ces méthodes. La segmentation en une suite de mots procède par un découpage arbitraire ; elle laisse de côté les aspects syntaxiques et sémantiques du texte. Par conséquent, elle peut produire du bruit lors de la réponse à la requête de l'utilisateur.

La segmentation en phrases n'est pas fiable lorsqu'on attend en réponse une partie de texte ne nécessitant pas de travail d'inférence de la part de l'utilisateur, sachant que la phrase ne présente pas de garantie de complétude syntaxique. De la même façon que la segmentation en phrases, celle en paragraphes n'est pas non plus suffisamment fiable, du fait de la difficulté à interpréter un paragraphe dans des contextes dans lesquels il est rattaché à une unité qui le précède ou bien qui lui succède.

Les méthodes de segmentation thématique procèdent à l'identification des différents thèmes véhiculés par le texte, pour le segmenter en unités homogènes formant des blocs thématiques. Cette méthodologie se différencie de la nôtre, dans la mesure où nous cherchons à segmenter le document selon des critères de surface, pour procéder par la suite à son indexation. La première étape de notre travail consiste donc à chercher des délimiteurs pour segmenter le document en unités. Ces unités seront filtrées[1] pour ne garder que celles qui sont informatives, qui seront retenues pour l'indexation. Ainsi, nous adoptons une démarche inverse par rapport à la segmentation thématique.

Certes, la segmentation s'appuyant sur la structure logique reflétée par le sommaire présente l'avantage de donner en réponse des unités cohérentes. Cependant, dans notre problématique qui est de répondre à un besoin opérationnel, cette unité n'est pas suffisamment

[1] Le filtrage des unités segmentées ne fait pas l'objet de cet article.

fine et elle pourrait engendrer du bruit (Paganelli 97). Ce qui nous amène à nous investir dans une réflexion pour trouver une unité, qui pourrait être une unité de base pour l'indexation du document technique et qui pourrait satisfaire au maximum l'attente de l'utilisateur dans ce domaine.

4- Définition du modèle de segmentation

4.1 Caractéristiques de l'unité de segmentation

L'unité de segmentation, baptisée Unité Documentaire (UD), est une suite de formes (textuelles ou non textuelles), dont l'ensemble traite d'un point qui leur est commun. Les éléments entrant dans la composition de l'UD entretiennent des rapports de dépendance, telle que l'interprétation de l'un de ces éléments (E1) dépend de l'interprétation de ceux qui le précèdent (E-1, E-2, E-n) ou bien ceux qui lui succèdent. Autrement dit, chaque élément ne peut pas être compris en dehors de son ensemble constituant l'UD.

En effet, l'UD doit avoir une cohérence interne et une autonomie relative, exigeant toutefois une prudence pour ne pas conduire à l'isolement d'autres UD (autonomie linguistique de façon à pouvoir être lue et comprise indépendamment des unités adjacentes). Les éléments entrant dans la composition de l'UD sont liés au point que l'ensemble forme un tout relativement isolable du reste du document.

Ainsi, l'UD se construit en assurant une cohésion syntaxique et sémantique. Les principaux facteurs de cette cohésion sont :

- La co-référence thématique : tous les éléments de l'UD font référence à un même objet de discours (unité de contenu). L'UD doit assurer la perception d'une continuité, l'exploitation continue d'un même objet de discours[2].
- La compatibilité des propos : il faut que les éléments constitutifs de l'UD soient compatibles du point de vue de l'enchaînement logique des idées. On n'insère pas de marque d'UD devant une phrase qui représente la continuation des phrases précédentes. Il faut que les faits et les modalités mentionnés d'une phrase à l'autre ne soient pas contradictoires.
- Repérage facile en surface : dans un processus de traitement automatique, l'UD doit être repérée et caractérisée par des indicateurs formels présents en surface et non ambigus.

Pour définir l'UD à l'intérieur de la structure globale du document technique, nous tenons compte de l'ensemble des connaissances qui résident dans un document structuré : connaissances structurelle et linguistique. Nous formulons l'hypothèse que ces différents éléments jouent un rôle essentiel dans le marquage des UD. Ils peuvent ainsi donner des instructions de découpage du document en unités cohésives (Ouerfelli 01).

4.2 Principes du modèle de segmentation

Entre des approches de segmentation qui ne prennent pas du tout en compte la structure des documents (segmentation en une suite de mots) alors qu'il s'agit d'une caractéristique forte

[2] On utilise "discours" comme synonyme de "texte" dans le sens de Ricoeur (86) "le texte : tout discours fixé par l'écriture".

des documents techniques, et celles qui ne proposent que cela (la segmentation à partir du sommaire), nous devons parvenir à un équilibre.

Le statut de l'UD est intermédiaire entre la phrase et la plus petite unité logique répercutée dans le sommaire. Cette unité doit être délimitée grâce à des indicateurs de surface, pour cela il faut monter à un niveau plus haut que la phrase à l'intérieur de l'unité logique. Ce niveau correspond au paragraphe typographique (bloc de texte délimité par deux alinéas[3]). L'idée directrice de notre méthode de segmentation est de partir donc, d'une unité minimale (le paragraphe) pour chercher l'UD répondant aux propriétés requises (autonomie linguistique, cohésion syntaxique et sémantique), formant un bloc "thématique" homogène.

Pour définir notre modèle de segmentation, nous sommes partis de la réalité des documents techniques pour en induire des critères de segmentation. En ce sens, nous avons mené une étude de 5 manuels techniques.

4.2.1 Etude du corpus

Les manuels techniques couvrent trois domaines différents (électronique, aéronautique et l'informatique). Nous avons essayé de voir le statut du paragraphe dans ces manuels, sachant que nous avons défini 5 types de paragraphes dans le document technique, en plus des objets non textuels (tableau et figure). Ces paragraphes sont :

- *Paragraphe titre présent ou non dans le sommaire* : il se distingue des autres types de paragraphes par sa forme typographique (mis en relief) et sa fonction comme indice de rupture avec ce qui précède et indice de continuité avec au moins le paragraphe qui lui succède directement.
- *Liste d'éléments* : elle est formée d'un ensemble d'éléments (items) dont l'alinéa constitue l'ouverture et la fermeture de cet ensemble.
- *Paragraphe textuel* : il se compose d'un nombre variable de phrases, qui sont souvent autonomes. Le critère formel caractérisant ce paragraphe est son ouverture par une majuscule et sa fermeture par un signe de ponctuation forte et surtout son encadrement par deux alinéas.
- *Paragraphe alphanumérique* : c'est une unité spécialisée, elle a pour fonction de présenter les paramètres d'exécution d'une application (énoncé de commandes, résultat d'exécution d'une commande etc.).
- *Paragraphe consigne* : il donne des instructions de mise en garde au lecteur pour l'exécution d'une tâche ou d'une procédure. Il est marqué par des indicateurs lexicaux bien définis du genre (Attention, Remarque, Important etc.).

4.2.2 Résultats

Les principaux résultats dégagés de cette étude sont :

Paragraphe titre :

- *Présent dans le sommaire* **:** il ne sera pas retenu pour la segmentation, il servira comme une hiérarchie lors de la réponse à la requête de l'utilisateur par la suite[4], sauf dans les cas où il est suivi : d'un paragraphe textuel débutant par un marqueur de

[3] L'alinéa est défini comme un retour à la ligne suivi d'un saut de ligne et / ou d'un retrait.

[4] Les titres présents dans le sommaire constituent des niveaux génériques. De ce fait, il serait intéressant de les dissocier des paragraphes adjacents et ne pas les considérer dans la segmentation. Ils constitueront l'environnement hiérarchique de la réponse qui se trouvera ainsi structurellement encadrée et gagnera en pertinence et en lisibilité.

continuité[5], d'une liste d'éléments ou d'un paragraphe alphanumérique. Dans ces cas, le titre doit s'afficher en-tête de ces paragraphes pour garantir une certaine autonomie linguistique et une cohésion sémantique, critères primordiaux pour la constitution de l'UD.

? *N'est pas présent dans le sommaire :* il sera retenu pour la segmentation. Il sera un composant de l'UD formée par l'objet textuel ou non textuel qui le suit immédiatement.

Liste d'éléments : elle s'inscrit souvent dans la continuité du ou des paragraphe(s) adjacent(s) (avant et / ou après) du genre paragraphe titre ou paragraphe textuel.

Paragraphe textuel : il pourrait constituer une UD autonome à condition qu'il ne débute pas ou ne finisse pas par l'un des marqueurs de continuité. Il pourrait aussi se combiner avec un autre paragraphe du même type pour constituer une UD.

Paragraphe alphanumérique : il s'inscrit dans la continuité du ou des paragraphe(s) adjacent(s) (avant et / ou après) du genre paragraphe titre ou paragraphe textuel.

Paragraphe consigne : il ne bénéficie d'aucune autonomie, il est toujours lié à ce qui le précède. Cependant, dans certains cas il peut être rattaché à deux ou plusieurs UD ce qui pose le problème de son lien avec l'entourage textuel dont il fait partie. De ce fait, il sera lié au paragraphe qui le précède immédiatement pour constituer une UD. Il sera aussi stocké dans un fichier à part pour qu'il soit rattaché aux différentes UD figurant dans la même unité logique indexée par le sommaire.

Tableau : c'est un élément qui est généralement associé au(x) paragraphe(s) qui le précède(nt). Il apporte de l'information complémentaire pour ce(s) paragraphe(s). Ce type d'objet n'est pas autonome pour constituer une UD.

Figure : cet objet peut soit apporter de l'information explicative du texte qui le précède ; soit interpréter le contenu de ce texte lorsqu'il est un peu complexe. Par conséquent, la figure est toujours liée au(x) paragraphe(s) la précédant. Elle ne peut pas constituer une UD autonome. Il est à noter que pour le traitement des figures se pose le problème de lien entre paragraphes textuels et figures discontinus[6].. Pour essayer de résoudre ce problème, nous pensons utile de segmenter le document en tenant compte de cet objet, mais aussi de stocker les figures dans un fichier et on aura recours à ce fichier lors de la réponse à la requête de l'utilisateur, si le paragraphe dans lequel se trouve le renvoi constitue une réponse à sa question.

Voilà donc un bilan général des résultats recueillis à partir de l'étude de notre corpus. Ces résultats ont été confirmés par 14 juges qui ont participé à l'évaluation de notre modèle de segmentation. La méthodologie adoptée consiste à confirmer la validité cognitive de notre modèle de segmentation. Pour cela, nous avons procédé à l'évaluation de 14 extraits segmentés en deux UD adjacentes (sept extraits non ambigus et 7 extraits ambigus). La segmentation des extraits est jugée non ambiguë lorsqu'elle répond aux propriétés données à l'UD. En revanche, la segmentation des extraits ambigus ne répond pas aux propriétés de l'UD. Cette segmentation pose le problème de traitement de certains paragraphes (titre, consigne) et celui de la rupture entre deux UD adjacentes en l'absence de marqueurs de continuité qui sont pourtant sémantiquement liés. Les extraits ambigus sont choisis intentionnellement pour mettre à l'épreuve les lacunes de notre modèle de segmentation.

[5] Nous reviendrons sur les marqueurs de continuité plus loin.
[6] Ce lien s'exprime souvent par un renvoi dans le paragraphe textuel vers une figure présente dans un autre endroit dans le document.

A partir de l'étude du corpus et de la validation psychologique, nous avons essayé d'établir un certain nombre de règles de segmentation.

4.3 Propriétés des règles de segmentation

Les règles de segmentation en UD reposent sur les marqueurs de continuité et la nature des relations existantes entre les constituants des UD.

4.3.1 Les marqueurs de continuité

Les marqueurs de continuité[7] permettent de lier les objets (textuels ou non textuels) adjacents ensemble grâce à des indicateurs de surface de nature linguistique ou typographique :

Marqueurs linguistiques

- *Marqueurs d'Intégration Linéaire (MIL) : "*si, alors, ensuite, Aussi, De plus, d'autre part" au début du paragraphe textuel[8].
- *Mots de liaison :* "par exemple, pour cela, pour ce faire, par ailleurs, de tel cas" au début du paragraphe textuel.
- *Reprise anaphorique :*

  Démonstratif : "ce, cette, ces, ceci" au début du paragraphe textuel.

  Pronom personnel : "il, elle" au début du paragraphe textuel, à l'exception des tournures impersonnelles du genre (*il convient de rappeler* ; *il est à noter* etc.).
- *Marqueurs de renvoi avant et arrière :* "ci-dessous, comme suit" pour les renvois avant. "Ci-dessus" pour les renvois arrière. Ces marqueurs apparaissent au début ou à la fin du paragraphe textuel.
- *Autres connecteurs :* "également, en particulier" au début du paragraphe textuel.

Marqueurs de ponctuation : l*es deux points (:)* à la fin du paragraphe textuel.

4.3.2 Typologie de relations entre les constituants de l'UD

Les résultats dégagés suite à l'étude du corpus et de la validation, nous inspirent une typologie de relations entre les différents objets textuels et non textuels susceptibles de figurer dans les UD.

Ces relations sont réparties en : relation descendante, relation ascendante et relation ascendante / descendante.

Relation descendante : cette relation concerne le paragraphe titre indexé ou non par le sommaire. Il est toujours lié à ce qui le suit (un ou plusieurs paragraphes textuels, liste d'éléments, etc.). On peut le considérer en plus de son statut de paragraphe, comme opérateur de liaison.

Relation ascendante : cette relation concerne les éléments qui sont inscrits dans la continuité de ce qui précède dans le texte. Elle s'applique donc sur le paragraphe consigne et la figure.

Relation ascendante / descendante : ce type de relation concerne le paragraphe textuel, la liste d'éléments, le paragraphe alphanumérique et le tableau.

---

[7] Les fonctionnalités de ces marqueurs sont discutées longuement dans Ouerfelli (01).
[8] Soit en-tête du paragraphe soit à l'intérieur de la première phrase.

Pour le paragraphe textuel, la relation ascendante peut s'exprimer par un marqueur de continuité de nature linguistique au début du paragraphe (reprise anaphorique, mot de liaison etc.). En revanche, la relation descendante peut s'exprimer par les deux points à la fin du paragraphe.

La liste d'éléments, le paragraphe alphanumérique et le tableau ne peuvent pas être isolés de leur contexte. Ils ont une relation ascendante avec ce qui les précède (paragraphe textuel, paragraphe titre). Ils peuvent avoir aussi une relation descendante avec ce qui leur succède, généralement un paragraphe textuel marqué par un indice de continuité au début.

Il est important de préciser ici que la relation ascendante / descendante n'est pas double, mais elle est indéterminée ; on peut trouver une liste d'éléments qui est rattaché à un paragraphe qui la précède sans l'être à celui qui lui succède. Aussi, on peut trouver un paragraphe textuel lié à un autre paragraphe qui lui succède sans l'être avec celui qui le précède.

Il est à noter aussi que ces différentes relations sont toujours identifiées à partir de l'étude de notre corpus. On constate ainsi, le rôle des marqueurs de continuité dans ces relations. C'est en fonction de ces marqueurs que seront établies les règles de construction des UD.

4.3.3 Les règles de segmentation

Les constituants de chaque UD forment une séquence linéaire P1, P2, Pi, Pn. Ils seront étiquetés de la manière suivante :

TIS pour le paragraphe titre indexé par le sommaire.

TNIS pour le paragraphe titre non indexé par le sommaire.

PT pour le paragraphe textuel.

PA pour le paragraphe alphanumérique.

LE pour la liste d'éléments.

PC pour le paragraphe consigne.

ONT pour l'objet non textuel (tableau, figure).

Chaque constituant de l'UD peut, soit avoir un marqueur de continuité avant (M av) ou après (M ap), soit ne pas avoir aucun.

Dans ce qui suit, nous présentons la liste des règles définies :

Si *Pi est TIS* **Alors** *stocker dans le fichier titre*[9].

Si *Pi est TNIS* **Alors** *ne pas couper entre Pi et Pi + 1.*

Si *Pi est LE* **Alors** *ne pas couper entre Pi - 1 et Pi.*

Si *Pi est PT + M av* **Alors** *ne pas couper entre Pi - 1 et Pi.*

Si *Pi est PT + M ap* **Alors** *ne pas couper entre Pi et Pi +1.*

Si *Pi est PA* **Alors** *ne pas couper entre Pi - 1 et Pi.*

Si *Pi est PC* **Alors** *stocker dans le fichier consigne*[10].

et *ne pas couper entre Pi -1 et Pi.*

[9] Les titres dans ce fichier serviront à inscrire les UI dans leur hiérarchie logique.
[10] Ce fichier servira à lier le paragraphe consigne aux différentes UD figurant dans la même unité logique.

Si *Pi est ONT* **Alors** *ne pas couper entre Pi - 1 et Pi.*

Si *Pi est ONT = "figure"* **Alors** *stocker dans le fichier figure*[11].

Ce qui implique Pi =

Stockage si Pi est TIS, PC ou ONT = "figure"

Ne pas couper de Pi - 1 si

Pi est LE

ou Pi est PT + M av

ou Pi est PA

ou Pi est PC

ou Pi est ONT.

Ne pas couper de Pi + 1 si

Pi est TNIS

ou Pi est PT + M ap.

Sinon couper.

Entre ***Pi et Pi + 1***

Et

Entre ***Pi et Pi - 1***

Les règles de segmentation reflètent les différents cas d'empaquetage des paragraphes et objets non textuels ensemble recensés dans l'étude du corpus. Pour cela, nous avons pris appui sur les marqueurs de continuité pour élaborer œs règles. Ainsi, on peut se demander sur le rôle exacte de ces règles : pour regrouper les paragraphes ensemble ou bien pour segmenter comme nous le préconisons.

En effet, le rôle de ces règles est de regrouper les paragraphes à partir des critères de surface pour construire des UD. Ces règles établiront les frontières entre les paragraphes adjacents selon les propriétés données à l'UD. La frontière se fait entre les paragraphes en l'absence de marqueurs de continuité, ce qui implique en principe un changement du point de vue[12].. Cette frontière correspond donc à la fermeture d'une UD et l'ouverture d'une autre UD. De ce fait, les règles définies permettront de segmenter le document en UD.

Il est intéressant de noter que ces règles ne concernent pas les cas particuliers[13] correspondant au lien inter paragraphes en l'absence de marqueurs de continuité (deux paragraphes adjacents qui sont sémantiquement liés sans présence de marqueurs de continuité, ou bien le lien entre un paragraphe et une figure discontinus).

[11] Ce fichier servira à résoudre le problème des liens entre objet textuel et figure discontinus.

[12] L'absence de continuité entre deux paragraphes adjacents et en l'occurrence l'absence d'un marqueur de continuité signifie un changement dans le point de vue sous lequel est envisagé le sujet. Ce qui constitue une raison suffisante pour introduire une nouvelle UD.

[13] Ces cas ne font pas l'objet de cet article. Ils sont discutés dans Ouerfelli (01).

4.3.4 Exemple d'application des règles de segmentation

Paragraphe textuel (PT)

Si *Pi est PT + M av* ***alors*** *ne pas couper entre Pi et Pi - 1*

Si *Pi est PT + M ap* ***alors*** *ne pas couper entre Pi et Pi + 1*

Sinon *COUPER*

entre Pi et Pi -1

et

Pi et Pi +1

Segmentation d'un extrait en fonction des règles précédentes :

(...)

Les calculs de structures pour la configuration de décollage et d'atterrissage doivent être basés sur la limite élastique du matériel (...) Les articles d'expérience en free- floating doivent être conçus pour tenir une charge éventuelle de 2, 5g dans toutes les directions en raison de la possibilité de choc sur une extrémité ou une paroi à la suite d'une manoeuvre.

¶

Chaque analyse de structure doit comprendre au minimum :

¶

1. Plan ou schéma de la structure.

2. Résultats des calculs de contraintes.

3. Masse des composants.

4. Propriétés des matériaux.

¶

Les calculs de structures doivent être joints à la description des expériences (...)

5- Conclusion et perspectives

Le traitement du document technique en amont de l'indexation s'effectue en deux étapes complémentaires : segmentation en UD et filtrage des Unités Informatives (UI). C'est sur la qualité de la première étape que repose l'efficacité de l'indexation. En effet, si l'on découpe en unités trop courtes, l'information risque d'être dispersée, ce qui conduit l'utilisateur à faire beaucoup de chemin pour trouver une réponse à sa question. A l'inverse, des unités trop longues risquent de faire perdre des degrés de pertinence, dans la mesure où l'information recherchée peut être noyée. Ainsi, il faut éviter ces deux problèmes, ce que nous avons essayé de faire en optant pour une segmentation non pas *a priori* (en fonction d'une unité fixe définie à l'avance), mais plutôt dynamique dont le paragraphe est l'unité de base. Cela présente l'apport principal de notre travail, pour lequel nous sommes partis de la réalité des documents techniques pour définir des critères de segmentation.

Le résultat de notre travail est un ensemble de règles de segmentation. A partir de ces règles, nous pouvons établir un modèle formel pour l’implémenter dans une optique de résolution automatique de la segmentation des documents techniques dont la fiabilité est réelle.

Bibliographie

[Bronckart 85] J. P. Bronckart et al.- *Le fonctionnement du discours*.- Neuchâtel ; Paris : Delachaux & Niestlé Editeurs, 1985.- 175p.

[Catach 94] N. Catach.- *La ponctuation*.- Paris : PUF, 1994.-127p.

[Charolles 88] M. Charolles.- Les plans dorganisation textuelle, périodes, chaînes, portées et séquences.- in : *Pratiques*, n°57, mars 1988. - pp. 3-13.

[Corblin 95] F. Corblin.- *Les formes de reprise dans le discours : anaphores et chaînes de référence*.- Rennes : PUR, 1995.- 246p.

[Drillon 91] J. Drillon.- *Traité de la ponctuation française*.- Paris : Gallimard, 1991.- 469p.

[Fourrel 97] F. Fourel.- Impact de la structure du document sur la recherche d'informations.- in : *Ingénierie des systèmes d'informations*, vol. 5, n°3, 1997.- pp. 339-365.

[Froissart 99] C. Froissart, G. Lallich.- Document technique : unicité et pluralité.- in : *journées ISKO*, Lyon, 21-22 octobre 1999.

[Le Coadic 98] Y. F. Le Coadic, *Le besoin d'information : formulation, négociation, diagnostic*, Paris : ADBS, 1998, 191p.

[Luc 99] C. Luc, M. Mojahid, J. Virbel.- Connaissances structurelles et modèles nécessaires à la génération de textes formatés.- in : *2 ème colloque francophone en Génération Automatique de Textes*, Grenoble, 30 septembre - 1 er octobre 1999.- pp. 157 - 169.

[Mochizuki 98] H. Mochizuki, T. Honda et M. Okumura.- Text segmentation with multiple surface linguistic cues.- in : *COLING-ACL*, 1998, pp. 881-885.

[Ouerfelli 01] T. Ouerfelli, *La segmentation des documents techniques composites dans une perspective d'indexation : vers la définition d'un modèle dans une optique d'automatisation*, Thèse de Doctorat en sciences de l'information et de la communication, Université Stendhal de Grenoble 3, 2001, 223p.

[Ouerfelli 99] T. Ouerfelli, G. Lallich, Base textuelle structurée et indexation : l'exemple de la documentation technique, *Colloque international en sciences de l'information*, Tunis 3 - 5 mars 1999.

[Paganelli 97] C. Paganelli, *La recherche d'information dans des bases de documents techniques en texte intégral. Etude de l'activité des utilisateurs*, Thèse de Doctorat en sciences de l'information et de la communication, Université Stendhal de Grenoble 3, 1997, 354p.

[Salton] G. Salton, A. Singhal, C. Buckeley, Automatic text decomposition and text themes, *7 th conference in Hypertext*, Washington, march 1996, pp. 53 - 65.

[Turco 88] G. Turco, D. Coltier.- Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire.- in : *Pratiques*, n°57, mars 1988.- pp. 57-79.

[Vigner 76] G. Vigner, A. Martin.- *Le français technique*.- Paris : Hachette, 1976.- 111p.

[Zobel 95] J. Zobel, A. Moffat, R. Wilkinson.- Effecient retrieval of partial documents.- in : *Information processing & management*, vol. 31, n°3, 1995.- pp. 361-375.