

Pratiques d' indexation dans les Bases Textuelles Structurées : Application aux Textes Techniques sous Format HTML

Tarek Ouerfelli et Geneviève Lallich-Boidin

Université Stendhal

Résumé

Parvenir à un système de recherche d' information technique capable de donner en réponse une unité de texte pertinente nécessite en amont de l' indexation une phase de segmentation du texte en Unités Documentaires fines et cohérentes. Dans cet article, nous avançons quelques éléments sur la stratégie de segmentation qui pourra être adoptée dans un processus d'indexation d'un texte technique volumineux. La stratégie de segmentation adoptée sera présentée à partir des résultats d'une expérimentation menée sur plusieurs textes techniques sous format HTML. Cette stratégie se veut dynamique avec la prise en compte des différents médias qui existent dans les textes techniques (texte, figure, tableau,). Ainsi, l'Unité Documentaire à indexer, qui se trouve à l'intérieur de l'unité logique peut être une unité textuelle composée d'un ou de plusieurs paragraphes ou bien une unité composite de paragraphe, tableau et figure.

1. Introduction

La recherche documentaire classique interroge des bases documentaires dont les enregistrements sont des références de documents. Ainsi, la problématique de la recherche d' information dans ces systèmes consiste à partir de l' expression de la requête d' un utilisateur, qui définit le thème de la recherche, à retrouver un ensemble de références de documents dont le contenu correspond au thème recherché, parmi l' ensemble des documents de la base documentaire considérée.

Les progrès technologiques (stockage magnétique, édition structurée) permettent maintenant de disposer des documents intégraux sur support informatique. L'information stockée dans la base peut comporter à la fois la description du document (physique et intellectuelle) et le texte intégral du document. L'avènement de ces technologies a fait apparaître une nouvelle génération de bases de données, à savoir les bases de données textuelles, parmi lesquelles on distingue : *les bases de documents textuels* qui contiennent des textes de documents, nombreux, peu volumineux et peu structurés (articles de presse) et les *bases textuelles structurées* qui contiennent le texte d'un ou de plusieurs documents, de forte structuration et de taille volumineuse. Une application de cette dernière catégorie est la recherche d'information dans les manuels techniques de description, et d'utilisation de dispositifs techniques complexes, comme celles de gros logiciels ou d'appareillages complexes. C'est dans ce cadre, que nous situons le présent travail. L'indexation de ces documents soulève des problèmes spécifiques liés aux spécificités des textes techniques (hétérogénéité, taille volumineuse et l'application à visée opératoire de la recherche d'information dans ces textes). Cette tâche se réalise en deux étapes complémentaires :

- Découpage du texte en Unités Documentaires ¹(UD) qui seront utilisées comme base pour son indexation.
- Représentation du contenu de ces UD.

La première étape constitue l'axe principal de ce travail. Ainsi, nous présentons dans un premier temps les propriétés spécifiques du texte technique. Cette présentation nous conduit à définir dans un second temps, les pratiques d'usage et l'indexation de ce type de texte. Enfin, nous avançons quelques éléments sur la stratégie de segmentation, qui pourra être adoptée dans un processus d'indexation d'un texte technique volumineux.

2. Propriétés des textes techniques

Nous désignons par texte technique, un document de type manuel d'utilisation de dispositifs techniques. Il est fortement structuré avec une organisation logique

bien définie. Le texte technique véhicule des savoirs et des savoir-faire propres à un champ technique particulier. Il représente aussi bien la description d'une machine (avion, train, radar, 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. De surcroît, le texte technique s' adresse à un groupe socialement et professionnellement homogène d' utilisateurs, qui effectuent une activité opératoire et partagent des connaissances communes. Il vise à informer non pas un individu particulier mais un élément indifférencié de cette catégorie socio-professionnelle, de la réalité du fonctionnement, et des propriétés d' un dispositif technique.

Le texte technique se caractérise par un certain nombre de caractéristiques, dont principalement l'hétérogénéité. Il peut comporter des informations sous forme différente (texte, image, figure, tableau) ; il peut comporter aussi différentes typologies d'unités textuelles du genre premier plan / second plan (Caro 95) ou bien objet / action (Paganelli 97). La représentation de son contenu doit prendre en considération ces caractéristiques, ainsi que, l'application visée par son utilisation, en l'occurrence l'objet de la recherche d'information dans ce texte, qui est de nature opératoire. C'est essentiellement ce qui différencie les textes techniques des autres types de textes.

3. Pratiques d'usage des textes techniques

La recherche d' information dans les textes techniques est essentiellement à visée opératoire, pour réaliser une tâche ou une action (Vigner 76, Bronckart 85). Un utilisateur expert recherche de l' information en vue de répondre à un besoin professionnel. Lors de la consultation d'un manuel technique, l'utilisateur effectue une recherche dans le but de savoir pour faire. Il cherche généralement à atteindre directement une donnée précise, l'information la plus élémentaire satisfaisant son besoin. Ainsi, le processus de recherche d'information se doit alors d'être particulièrement rapide et efficace, d'où vient l'intérêt de traiter le texte technique comme une construction moléculaire susceptible d'être

décomposée en unités plus fines, pour donner naissance à de nouvelles UD's utilisables, 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. Ainsi, nous devons tenir compte de ces différents aspects, pour l'indexation du texte technique. Cette opération qui doit se faire non plus sur le texte dans sa globalité, mais sur des parties autonomes du point de vue "sémantique" et syntaxique. Ainsi, ce type de texte présente des propriétés d'usage spécifiques qui exigent leur prise en compte dans le cadre de la conception d'un système automatisé de recherche d'information, pour garantir une certaine efficacité de ce système dans la réponse aux demandes de ses utilisateurs.

Parvenir à un système de recherche d'information technique capable de donner en réponse une unité de texte pertinente nécessite donc, en amont de l'indexation une phase de segmentation en unités fines et cohérentes. Cette question représente la problématique soulevée dans notre travail de recherche mené à l'équipe CRISTAL², sur laquelle nous allons apporter des éléments de réponse.

4. La segmentation pour l'indexation

Un certain nombre de travaux en recherche d'information se sont intéressés à la segmentation des textes avec deux préoccupations : la première répond au souci de ne pas noyer l'utilisateur sous une masse trop importante de documents en réponse à une requête lorsque les documents sont de taille importante. La seconde réside dans une plus grande efficacité de la recherche elle-même. Ainsi, au lieu de calculer une similarité entre une requête et la totalité d'un texte, on mesure la similarité entre cette requête et chacune des UD's constituant le texte. Ceci va tout à fait dans le sens de notre objectif de départ, à savoir la segmentation du texte technique pour un accès plus précis et plus localisé à l'information à base d'UD's fines à l'intérieur de la structure globale du texte. 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 un processus de recherche d'information, on distingue différentes méthodes de segmentation :

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

Ces différentes méthodes sont présentées en détail dans Lallich & Ouerfelli (98). 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.

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 fine et elle pourrait engendrer du bruit lors de la réponse à la requête de l'utilisateur. Ce qui nous amène à nous investir dans une réflexion pour trouver une UD qui pourrait être une unité de base pour l'indexation du texte technique et qui pourrait satisfaire au maximum l'attente de l'utilisateur dans ce domaine.

5. Stratégie de segmentation dynamique

5.1 Principe

Entre des approches 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 de nos textes techniques, et celles qui ne proposent que cela (la segmentation à partir du sommaire), nous devons parvenir à un équilibre. La problématique qui se pose dans cet ordre d'idées est celle de l'unité à choisir pour découper le texte technique, qui servira par la suite comme base pour son indexation.

Nous souhaiterons définir cette unité par une méthodologie qui requiert la prise en compte de l'ensemble des connaissances qui résident dans un texte structuré : connaissance structurelle, dispositionnelle et typographique. 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 texte en unités cohésives (Ouerfelli & Lallich 99).

Le statut de l'UD est intermédiaire entre la phrase et la plus petite unité logique répercutée dans le sommaire. Cette unité doit être repérée, délimitée et caractérisée grâce à des indicateurs formels présents en surface et non ambigus. Ainsi, il faut monter à un niveau plus haut à l'intérieur de l'unité logique, ce niveau correspond au paragraphe typographique. Ainsi, l'idée directrice de notre méthode de segmentation est de partir d'une unité minimale ici le paragraphe typographique pour chercher l'UD répondant aux propriétés requises (autonomie linguistique, cohésion syntaxique et sémantique) formant un bloc "thématique" homogène (Ouerfelli & Lallich 99).

5.2 Segmentation en paragraphes

Dans cette section, on va présenter les résultats d'une expérimentation faite sur trois manuels techniques sous format HTML. La finalité de cette expérimentation est de détecter les marqueurs de changement d'une UD à une autre permettant de donner des critères de segmentation du texte technique. Avant de présenter ces résultats on va par la suite définir quelques termes essentiels dans notre étude pour préciser la manière dont chaque terme est utilisé.

5.3 Définition des termes

Paragraphe et alinéa

Par commodité, nous désignons dans ce travail par *paragraphe* : un bloc de texte délimité par deux *alinéas*.

L'alinéa dans notre travail est considéré comme un signe typographique jouant le rôle de délimiteur de paragraphe. A priori l'alinéa correspond suivant le cas, à un retour à la ligne
 suivi d'une indentation <BLOCKQUOTE> et ou d'un saut de deux lignes <P>. Cette idée rejoint l'avis de Bessonnat (88) qui considère l'alinéa comme un signe blanc qui délimite une quantité voire le texte du paragraphe.

Typage des paragraphes

Le bloc de texte délimité par les deux alinéas peut avoir différentes valeurs et en l'occurrence différents critères de formes. Ainsi, on distingue différents types de paragraphes dans les textes techniques. Chaque paragraphe se caractérise par sa fonction et sa forme :

- *Paragraphe titre* : est un paragraphe particulier. 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* : la liste est un objet textuel que l'on peut considérer comme un bloc cohésif qui vise à transmettre une information d'une façon bien précise. Généralement, cet objet textuel répond à une grammaire (annonce de liste + items + clôture). Dans ce travail, nous allons traiter les listes comme des objets textuels formés d'un ensemble d'éléments (items) dont l'alinéa constitue l'ouverture et la fermeture de cet ensemble. Ces listes peuvent être ordonnées dans un ordre numérique (1, 2, 3 ...) ou bien alphabétique (a, b, c, ...). Ce type de liste est inscrit entre les deux balises d'ouverture et de fermeture et . Elles peuvent aussi être marquées par des puces (•) entre les balises et .
- *Enoncés de commandes* : ce sont des unités spécialisées, elles ont pour fonction de présenter les paramètres d'exécution d'une application dans le langage du domaine qui est compréhensible par la machine. Elles ont généralement des

caractéristiques spéciales du point de vue typographique (caractères et polices se distinguent du texte principal) et de mise en valeur qui facilitent leur repérage du genre :

<code>lg_INTERVAL *(<VALEUR> *(LIGNE *(<<VALEUR> ") <element1> <element2></code>

- *Paragraphe textuel* : nous rangeons dans ce type tout paragraphe ne correspondant pas aux différents types évoqués ci-dessus. Le paragraphe textuel se compose d'un nombre variable de phrases qui sont souvent autonomes du point de vue syntaxique et sémantique. On distingue deux catégories de paragraphe textuel : le paragraphe textuel au sens classique communément admis, correspondant aux paragraphes que l'on trouve dans tout texte. Le critère formel caractérisant ce paragraphe est son ouverture par une majuscule et sa fermeture par un signe de ponctuation souvent une ponctuation forte et il est encadré entre deux alinéas.

La seconde catégorie correspond au "*paragraphe consigne*" : donnant des instructions au lecteur pour l'exécution d'une tâche ou d'une procédure. Le paragraphe consigne se caractérise par des critères formels de nature typographique (gras, italique) et il est généralement marqué par des indicateurs lexicaux bien définis (du genre *remarque, important, attention, recommandation,...*), qui expliquent sa fonction spécifique, et permettent de le distinguer du texte principal.

6. Expérimentation

6.1 Matériel linguistique étudié

Nous avons choisi donc de nous attaquer à un corpus technique composé de 3 manuels d'utilisateurs dans des domaines techniques variés. Les 3 manuels sont en langue française avec une organisation logique très forte. Ces manuels sont :

- "*Campagne de vols paraboliques : manuel des utilisateurs*": ce manuel décrit les techniques du vol parabolique : technique de pilotage, la manœuvre parabolique, etc.

- **"XCOMPIL : Guide d'utilisation"** : ce manuel présente les caractéristiques d'un outil "XCOMPIL" pour compiler les fichiers informatiques.
- **"Manuel d'utilisation de l'Analyseur Logique Tektronix 1241"** : il décrit les propriétés d'un analyseur de l'état électrique des circuits à microprocesseurs.

Les différents manuels traitent des domaines différents (aéronautique, informatique ou bien l'électronique). Cette diversité dans le contenu et la forme devrait permettre d'aboutir à des critères de segmentation applicables sur des corpus techniques de nature différente. Ainsi, ce choix de travailler sur plusieurs textes techniques nous permettrait de généraliser les règles de segmentation auxquelles nous aboutirons à la suite de ce travail d'investigation.

6.2 Traitement du corpus

6.2.1 Les objets textuels

Ces objets correspondent aux différents types de paragraphes définis supra.

a) Traitement des titres

Comme nous l'avons signalé, le paragraphe " titre " a un statut particulier par rapport aux autres types de paragraphes. On distingue deux types de titres dans le corpus traité :

- *Les titres non numérotés* : ils n'apparaissent pas dans le sommaire. Ces titres sont identifiés dans les documents sources par les balises <P>, et . La balise <P> concerne les sauts de deux lignes et la balise concerne la typographie des caractères qui est le gras. Ces titres sont souvent liés à l'unité textuelle qui les précède.
- Exemple du manuel des utilisateurs des vols paraboliques :

```
<P>
<B>Check-list des sources de risques </B>
<P>
<OL>
1. Matière, fluide inflammable et combustible (liquide, vapeur ou gaz) (...)
(...)
</OL>
<P>
```

Dans cet exemple, le paragraphe constitué d'une liste énumérative d'éléments présente un seul bloc. Intuitivement le titre est lié à ce bloc : il annonce cette liste et l'indexe en quelque sorte.

- *Les titres indexés par le sommaire* : ces titres sont numérotés dans le texte, ils représentent les différentes sections et sous-sections abordées dans le texte (H1, H2, H3, ...). Ils sont inscrits dans le sommaire et ils utilisent les mêmes critères typographiques que les titres non numérotés, mais avec des polices de caractères différentes en taille.

Ces titres constituent des niveaux génériques, de ce fait, ils ne seront pas retenus lors de l'indexation. Ce choix va tout à fait dans le sens de notre objectif de la méthode de segmentation visant à affiner la structure logique, c'est-à-dire chercher à l'intérieur de cette structure reflétée par les titres du sommaire, pour avoir un accès plus précis à l'information. Cependant, on ne procédera pas à leur élimination. Ils seront stockés dans un fichier à part et serviront comme hiérarchie lors de la réponse à la requête de l'utilisateur. Ils constituent l'environnement hiérarchique de la réponse. Ainsi, cette réponse se trouve structurellement cadrée, par conséquent, elle gagnera en pertinence et en lisibilité. Dans ce sens, Caro (95) rappelle que les titres et intertitres assurent une fonction de repérage lors de la lecture d'un texte. Ces titres sont considérés comme des balises permettant de naviguer dans l'univers informationnel du texte.

De ce fait, il nous semble nécessaire de présenter, en plus de l'unité réponse, le contexte dans lequel elle se situe. Ce qui permettra de la préciser et de l'enrichir. Ce principe ne s'appliquera que pour les titres correspondant aux UD's retenues pour l'indexation. Par conséquent, les titres des unités filtrées ne seront pas retenus, on fait allusion ici aux titres du genre " procédés " (Kerkouba 84) qui ne sont pas informatifs. Ils annoncent la fonction de l'unité textuelle qui suit (par exemple : présentation, introduction, conclusion).

b) Paragraphe Textuel

Dans la majorité des cas, les paragraphes constituant chaque section sont autonomes du point de vue syntaxique et sémantique. Dans ce sens, chaque paragraphe désigne un "thème" d'un point de vue particulier⁴.

Cependant, ce résultat n'est pas toujours valable sur tout le corpus ; on trouve, en effet, des paragraphes qui se suivent et qui sont intuitivement rattachés au niveau des idées véhiculées. De ce fait, pour assurer l'homologie entre la segmentation et l'articulation logique du texte, nous avons repéré les indices de surface qui ne mènent pas à l'insertion d'une marque d'UD entre deux paragraphes qui se suivent. Ceci va tout à fait dans le sens de notre conception de l'UD qui doit être considérée, non seulement comme une unité typographique, mais comme une unité de structure qui a des caractéristiques sémantiques et grammaticales. Elle doit présenter des critères de cohésion sémantique et syntaxique. Les indices repérés ne menant pas à un découpage en paragraphes seront présentés par la suite.

c) Paragraphe : liste d'éléments

Ces listes (à puces ou bien ordonnées) sont toujours liées à un morceau de texte précédent. Ce morceau peut être un paragraphe textuel ; généralement ce paragraphe se termine par le connecteur typographique (:) indiquant une ouverture vers un autre objet textuel. La liste peut être aussi précédée d'un titre avec des caractères typographiques (gras) différents du texte normal. Ces titres jouent le rôle, en quelque sorte, d'*indice de rupture* avec ce qui précède et d'*indice de continuité* avec la liste énumérative qui lui succède.

Les listes sont marquées aussi, par les sous-listes de différents niveaux (1, 2, 3). Chaque sous-liste est liée à un élément, soit de la liste principale (le cas de la sous-liste de niveau 1), soit de l'une des sous-listes (le cas des sous-listes de niveau 2 ou 3). Pour les différents niveaux des sous-listes, il faut toujours monter au niveau de l'élément de la liste principale, pour avoir une certaine autonomie et cohésion sémantique et syntaxique.

6.2.2 Les objets non textuels

Les objets non textuels dans le corpus correspondent aux tableaux et figures. Ces objets sont spéciaux et de ce fait, nécessitent un traitement particulier. Du point de vue typographique, les tableaux et les figures constituent des entités autonomes. Ils sont séparés des autres paragraphes par deux balises <P> correspondant à l'alinéa. Cependant, ces objets sont liés aux unités qui les précèdent : soit des paragraphes textuels ou bien un élément d'une liste. Il est à noter, qu'il existe deux types de tableaux :

- Les tableaux avec leur structuration en lignes et cellules dont la lecture est "convergente" entre lignes et cellules. Ils sont structurés par les deux balises <TABLE> et </TABLE>.
- Les tableaux avec leur structuration linéaire, dont la lecture est linéaire aussi. Ces tableaux sont présents dans le manuel des vols paraboliques. Ils sont marqués par les balises d'ouverture et de fermeture des listes à puces <TABLE> et </TABLE> .

Les différentes figures trouvées dans le corpus ont pour fonction, soit la *représentation* du contenu du paragraphe précédent, dans ce cas, elles apportent de l'information explicative sur l'objet textuel ; soit une fonction d'*interprétation* lorsqu'elles donnent des éléments pour l'interprétation du sens du paragraphe qui est un peu complexe. Dans ce sens, la figure apporte des informations complémentaires relatives à l'accomplissement du but du paragraphe. Ainsi, cette figure peut être considérée comme une unité de second plan selon la terminologie de Caro (95), de type para-texte. Elle joue le rôle de guidage de la compréhension, non plus de tout le texte, mais du paragraphe textuel ou bien du tableau qui lui est rattaché.

A partir des données évoquées ci-dessus, on peut conclure à la complémentarité entre l'objet textuel et l'objet non textuel dans le texte technique. D'une part, le tableau peut énumérer des objets signalés dans le paragraphe qui le précède, ainsi il apporte des informations complémentaires et nécessaire à l'interprétation du contenu informatif du paragraphe. D'autre part, le texte⁵ peut servir de contexte à la figure ; il peut contribuer ainsi à sa compréhension. De surcroît, la

figure peut remplir, par rapport au texte, une certaine fonction d'interprétation ou d'illustration de son contenu.

6.3 Regroupement des paragraphes textuels en Unités Documentaires textuelles

Il s'agit des cas où on décèle un enchaînement linéaire des idées entre les paragraphes textuels. Ainsi, on a repéré les marqueurs de continuité entre les paragraphes qui se succèdent. Ces marqueurs sont essentiellement de nature linguistique.

- Présence de Marqueurs d'Intégration Linéaire (*Si, Alors, Ensuite, Ainsi, De plus, également, d'autre part...*) au début du paragraphe.
- Présence des mots de liaison (*par exemple, Pour cela, Pour ce faire, ...*) au début du paragraphe.
- Reprise anaphorique au début du paragraphe :
 - a. Reprise par un démonstratif (*ce, cette, ces*)
 - b. Reprise par un pronom personnel (*il, elle*). Il est à noter que, les pronoms personnels se présentent rarement dans notre corpus, dans la mesure où, la description dans ces textes est une description d'objets et de tâches relatives à ces objets ou bien de procédures de réalisation d'une action dans un champ technique particulier et non pas de personnes. C'est le caractère impersonnel qui caractérise ces textes au sens de Vigner (76).

L'élément déterminant dans notre approche c'est que toute anaphore consiste en un lien entre deux segments textuels univoquement délimitables. Cette position conduit à considérer comme peu acceptable le cas où un pronom n'est pas précédé d'un antécédent. Par exemple, dans la phrase suivante extraite du manuel des utilisateurs des vols paraboliques (*Il est recommandé de planifier une campagne de vols paraboliques longtemps à l'avance,...*). La tournure est impersonnelle, de ce fait le pronom ne constitue pas un indice de continuité entre les deux blocs textuels

adjacents. Ainsi, l'antécédent⁶ se voit donc, attribuer un rôle central dans le mécanisme de continuité et le maintien de la cohésion, dans la mesure où, l'interprétation de l'anaphore exige l'emprunt d'un élément qui la précède fixant son interprétation. De cette façon, elle peut être considérée comme un indice de continuité entre des paragraphes textuels adjacents.

- Présence des marqueurs (*ci-dessous, ci-dessus*) : ces marqueurs pointent vers un objet textuel ou non textuel qui précède (ci-dessus) ou bien qui succède (ci-dessous). Ces indices sont souvent présents au début ou à la fin du paragraphe.

Voilà donc les différents éléments recensés dans le corpus jouant le rôle de relais et de continuité entre au moins deux paragraphes adjacents. Ces éléments constituent un dispositif de cohésion dans le texte. Il faut ajouter à ces éléments de nature linguistique le signe typographique exprimé par les (:). Ce signe de ponctuation, en marquant la limite d'un bloc, il sert à annoncer le suivant. Ce dernier apporte toujours une explication ou une précision (Védénina 89).

Il est à noter que, dans le corpus traité on a recensé des cas dans lesquels les unités adjacentes sont liées entre elles mais sans présence de marqueurs de surface permettant d'établir cette relation. De plus, on a recensé des cas dans lesquels les marqueurs de continuité ne remplissent pas leur fonction et sont ainsi, une source d'ambiguïté. Ces cas exigent un traitement particulier pour atteindre notre objectif tracé dès le départ à savoir une segmentation qui assure une homologie entre la pertinence des UD's et l'articulation logique du texte.

Une autre remarque concernant les paragraphes textuels du genre "*consigne*" qui ont un statut particulier. Ces paragraphes ne seront pas traités dans le présent article. Par contre, on a pas recensé dans le corpus des paragraphes "*énoncés de commande*" qui sont souvent inscrits dans la continuité d'une unité adjacente généralement un paragraphe textuel.

7. Conclusion

En terme de conclusion de cet article, on peut dire que le paragraphe peut servir comme une unité de base pour segmenter le texte technique en UD's tout en

tenant compte des paragraphes adjacents représentant un continuum dans le sens. Ce phénomène est représenté par les indices de continuité inter-paragraphique. La prise en compte de ces indices permet d'assurer une segmentation reflétant l'articulation logique des différents aspects traités dans le texte. Dans ce sens intervient le dynamisme de la méthode de segmentation adoptée, avec la prise en compte bien évidemment des différents médias qui existent dans le texte technique (texte, tableau, figure, ...). Ainsi, l'UD à indexer qui se trouve à l'intérieur de l'unité logique peut être une UD composée d'un ou de plusieurs paragraphes ou bien une unité composite de paragraphe, tableau et figure.

La phase de segmentation permet de découper le texte en plusieurs unités autonomes. Par la suite, il serait intéressant de voir s'il est envisageable de filtrer les unités, pour ne garder que celles qui sont informatives. Il serait important aussi de donner des critères de surface caractérisant ces unités comme non informatives. Cette perspective constitue la prochaine étape de notre travail de recherche.

Notes

¹ L'UD peut être composée *du texte seulement*, comportant généralement une ou plusieurs phrases. Elle peut être aussi *une combinaison du texte avec un tableau ou bien une figure* formant un ensemble relativement autonome et cohésif.

² Connaissances Recherche d'information Interfaces et Système de Traitement Automatique de la Langue.

³ HyperText Markup Language.

⁴ Le point de vue est l'angle ou l'aspect sous lequel est envisagé le sujet.

⁵ Le texte ici fait référence au paragraphe, l'objet de base de notre étude.

⁶ Ce qu'on appelle antécédent à savoir un segment de texte situé à l'avant de l'anaphorique dans le texte et ayant avec celui ci une relation d'interprétation ou de coréférence.

Bibliographie

- Bessonnat, D. 1988. Le découpage en paragraphes et ses fonctions. *Pratiques* 57 : 81-100.
- Bronckart, J. P. 1985. Le fonctionnement du discours, Neuchâtel ; Paris : Delachaux & Niestlé Editeurs.
- Caro, S. 1995. Rôle des organisateurs para-linguistiques dans la consultation des documents électroniques. *Thèse de Doctorat en Sciences de l'information et de la communication*, Grenoble 3.
- Kerkouba, D. 1984. Méthode d'indexation automatique des documents fondés sur l'exploitation de leurs propriétés structurelles. Application à un corpus technique. *Thèse de Doctorat en Informatique*, IMAG.
- Lallich, G. et T. Ouerfelli. 1998. La segmentation pour l'indexation d'un document technique : principe et méthodes. *Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique*, Sfax, 11-14 novembre 1998.
- Ouerfelli, T. et G. Lallich. 1999. Base textuelle structurée et indexation : l'exemple de la documentation technique. *Colloque international en sciences de l'information*, Tunis, 3-5 mars 1999.
- Paganelli, C. 1997 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*, Grenoble 3.
- Védénina, L. G. 1989. Pertinence linguistique de la présentation typographique, Paris : Peter.
- Vigner, G. 1976. Le français technique, Paris : Hachette.