Corela
Cognition, représentation, langage

3-2 | 2005
Vol. 3, n° 2

Structure matérielle et contenu sémantique du texte écrit

Marie-Paule Jacques

Édition électronique
URL : http://journals.openedition.org/corela/560
DOI : 10.4000/corela.560
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Marie-Paule Jacques, « Structure matérielle et contenu sémantique du texte écrit », *Corela* [En ligne], 3-2 | 2005, mis en ligne le 27 décembre 2005, consulté le 19 avril 2019. URL : http://journals.openedition.org/corela/560 ; DOI : 10.4000/corela.560

Structure matérielle et contenu sémantique du texte écrit

Marie-Paule Jacques

1. Introduction

1 Nous nous intéressons ici à l'articulation de la structure matérielle et du contenu sémantique de documents textuels écrits, en nous appuyant sur l'idée que la matérialité d'un texte participe à son sens.

2 A l'heure où les documents écrits se multiplient, un accès sélectif au contenu textuel devient un réel enjeu. Nul n'est désormais en mesure de lire intégralement tous les documents produits sur un sujet donné, il faut trier, cibler certains segments textuels selon le type d'information recherchée. Des outils toujours plus performants doivent permettre de faire face à l'inflation des écrits et à la nécessité d'un accès rapide à l'information.

3 La rapidité peut être obtenue par l'automatisation de la caractérisation du contenu d'un document, par l'extraction automatique de segments pertinents, c'est-à-dire des segments dans lesquels se trouve l'information pertinente pour représenter les idées et thématiques principales du document. Si on associe à cela un outil de visualisation dynamique[1], on offre à l'utilisateur le moyen de naviguer **sans se perdre** dans des documents longs tels que des rapports, des comptes-rendus de projet, des thèses, une documentation globale, etc., et d'emprunter divers chemins en fonction de ses besoins pour puiser l'information désirée.

4 Pour cela, il est nécessaire non seulement de pouvoir déterminer automatiquement **de quoi** les documents parlent, mais aussi de savoir **où** ils parlent de ce dont ils parlent, et encore **où** ils reviennent sur ce dont ils ont déjà parlé afin de mettre en évidence des continuités/contiguïtés ou ruptures/distances thématiques. Il s'agit de considérer le document comme constitué de segments organisés dans des parties, établissant éventuellement entre eux des liens à distance et des dépendances hiérarchiques,

permettant de **structurer** leur contenu. C'est par la prise en compte de cette structure que l'appréhension et la représentation du contenu du document peuvent être enrichies : la structure donne forme et sens au contenu, nous la concevons comme le fil d'Ariane d'une navigation sélective, c'est en la gardant en vue que le navigateur peut cheminer dans un document long au format numérique sans s'égarer.

5 Même si nous sommes guidée par des questions liées au traitement automatique, notre perspective est essentiellement et avant tout linguistique, notre étude se situe dans le cadre d'une linguistique de l'écrit. Nous nous proposons de décrire les propriétés des documents textuels écrits afin de déterminer comment des éléments spécifiques de l'écrit peuvent être exploités à des fins de traitement automatique.

6 Nous défendrons ici l'idée que la prise en compte de la structure matérielle – à la définition de laquelle nous consacrons le début de cet article – d'un document textuel est susceptible d'apporter une plus-value non négligeable pour la caractérisation de son/ses contenu(s), et ce parce que les deux sont indissolublement liés. Sur le plan de l'automatisation, il est d'autant plus pertinent de chercher à comprendre tout le parti que l'on peut tirer de la structure matérielle que celle-ci est relativement aisée à repérer dans un document numérique, surtout sous les formats qu'adoptent les textes conçus pour des supports numériques (html, sgml, xml...). Par *document numérique*, nous entendons aussi bien un document textuel disponible sous format numérique mais conçu pour être lu et/ou diffusé sous un autre format (papier ou autre), qu'un document textuel conçu pour être lu et/ou consulté sur un support numérique (par exemple, les pages du Web, les livres électroniques, etc.). Nous nous focalisons en tout premier lieu sur le premier type indiqué – document conçu pour le papier –, mais, comme ces deux types ne sont pas cloisonnés et que des passages du document conçu pour un support papier vers un format adapté à une consultation électronique sont possibles, nos propositions s'étendront au second type.

7 Les éléments de structure auxquels nous nous intéressons particulièrement ici sont les titres à l'intérieur des documents, c'est-à-dire les titres de sections et sous-sections. Nous montrerons (section 4.3) que ces éléments participent à la structuration des documents sur deux plans : sur le plan matériel, visuel, ils découpent et hiérarchisent le texte ; et sur le plan du contenu, ils peuvent jouer divers rôles qui vont de l'annonce d'une thématique à l'introduction de référents dans le discours, en passant par la focalisation sur des référents déjà présents.

8 Afin d'éclairer notre intérêt pour les titres, nous commencerons par exposer l'arrière-plan théorique sur lequel peut s'appuyer la prise en compte de la structure matérielle des textes écrits, après avoir défini celle-ci (section 2). Nous précisons de quelle façon les aspects liés à la matérialité du texte peuvent être exploités pour des tâches de traitement automatique, en donnant quelques exemples de recherches qui articulent le repérage de certains types d'informations dans le texte à la prise en compte de caractéristiques matérielles (section 3). Nous indiquons ensuite les premiers résultats de la recherche menée sur les fonctions des titres (section 4). Dans la section 5, nous revenons sur les perspectives de recherche à développer.

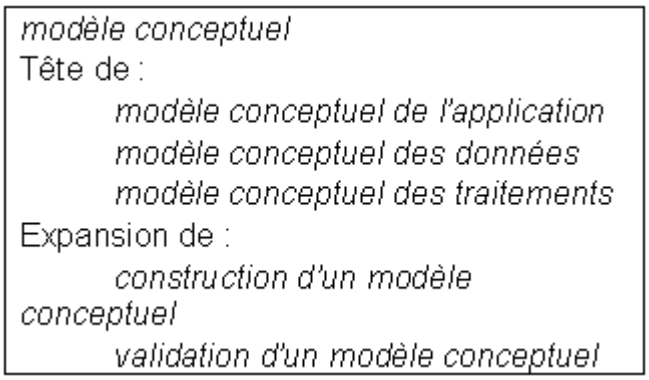
2. La structure matérielle des textes : arrière-plan théorique

2.1. Qu'est-ce que la structure matérielle des textes ?

9 Avant toute chose, il semble essentiel de préciser ce que nous entendons par *structure matérielle* des textes. S'agissant, comme nous l'avons indiqué précédemment, de documents écrits, c'est-à-dire conçus pour un support papier – et non de documents sonores, ou bien d'écrits qui seraient une transposition ou une transcription d'une forme orale –, les textes présentent des propriétés liées à leur support matériel. Une feuille de papier, de même qu'un écran d'ordinateur, de téléphone portable ou de livre électronique, offrent un certain **espace** sur lequel se dispose le texte. Cette propriété permet à l'espace de jouer un rôle dans la construction de la signification.

10 Pour illustrer ceci, arrêtons-nous sur une figure que nous empruntons à et qui, dans son texte d'origine, constitue un « Extrait du réseau terminologique construit par SYNTEX autour du candidat terme *modèle conceptuel* ».

Figure 1. Illustration du rôle de la matérialité dans la construction de la signification



Plusieurs éléments de la matérialité du texte participent à sa signification :

- l'alternance d'italique et de caractères droits permet d'attribuer des statuts différents aux éléments porteurs de ces caractéristiques ;
- l'indentation amène à mettre sur des plans différents le syntagme nominal *modèle conceptuel* et les autres syntagmes nominaux en italique ;
- les majuscules sur *Tête* et sur *Expansion* permettent de saisir ces deux expressions comme découpant des sous-ensembles (ce qui est redondant avec l'indentation).

11 On voit ici comment, à côté du lexique, l'inscription du texte sur un support spatial, en lui conférant des propriétés de mise en forme matérielle, apporte des éléments de signification. Pour produire le même contenu sémantique à l'oral, sans le secours des propriétés de forme et de disposition que la figure 1 exploite, il eût fallu une formulation beaucoup plus riche en éléments lexicaux, par exemple : « le terme *modèle conceptuel*[2] se retrouve comme Tête d'autres termes, les termes *modèle conceptuel de l'application*, *modèle conceptuel des données* et *modèle conceptuel des traitements* ; on le retrouve aussi comme Expansion des termes *construction d'un modèle conceptuel* et *validation d'un modèle conceptuel* ».

12 Il est parfaitement possible de trouver une formulation discursive pour une figure telle que la figure 1, mais on remarquera que cette formulation est bien moins économique que sa source écrite, et de ce fait il est à craindre que son traitement soit lourd sur le plan cognitif : le destinataire de la formule risque d'avoir oublié le premier terme mentionné avant même d'entendre le dernier – d'autant plus que les termes donnés en exemple présentent un certain niveau de complexité à la fois syntaxique et cognitive.

13 Dès lors que l'on a à traiter des séries, des ensembles ou des catégories, dès que l'on a à introduire des découpages et des hiérarchisations, l'écrit, de par ces propriétés de matérialité et de spatialisation que nous venons d'illustrer, s'avère particulièrement approprié. Et ce parce que l'écrit implique un traitement cognitif différent de celui requis par l'oral, précisément lié à sa matérialité.

14 Pour , l'écrit est un transformateur cognitif : le passage de la dimension linéaire de l'oral à la dimension spatiale de l'écrit implique une modification du traitement de la langue, en ce que l'introduction d'une dimension spatiale permet des manipulations formelles qui en retour transforment la pensée et les opérations mentales.

Afin de montrer ceci, autorisons-nous quelques petites manipulations de la figure 1 :

Figure 2. Transformation 1 de la figure 1

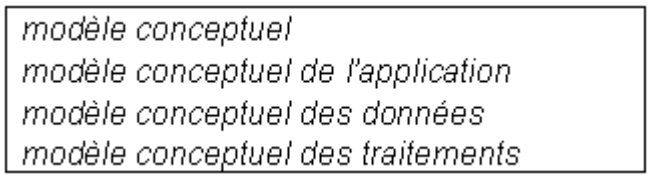
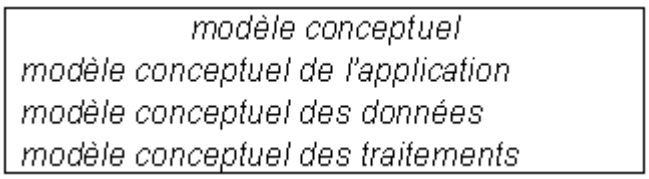


Figure 3. Transformation 2 de la figure 1



15 La figure 2 propose une liste de termes que la mise en forme matérielle met tous sur le même plan. Dans la figure 3, le simple déplacement du premier terme vers le centre suffit à lui accorder un statut spécifique par rapport aux autres éléments : on peut ainsi le comprendre comme un titre, comme le descripteur d'une catégorie formée des autres éléments, comme le plus petit dénominateur commun de ce qui suit, etc. De telles manipulations (dont nous ne donnons ici qu'un exemple très simplifié) ne sont pas sans incidence sur le plan conceptuel : selon la place physique qui lui sera accordée dans la liste, la première expression sera ou ne sera pas constituant d'une catégorie, ou bien en sera un exemplaire privilégié, etc. Appliquées sur une vaste échelle, les manipulations formelles permises par un support écrit sont effectivement de nature à modifier en profondeur les opérations mentales mises en jeu.

16 Ces quelques exemples mettent aussi en évidence le fait que l'organisation matérielle du texte n'est finalement que le reflet de son organisation logique. Là où l'auteur d'un document veut introduire une rupture, veut manifester une dépendance, veut exprimer une série, ou tout autre opération de structuration, il a à sa disposition, en plus des moyens lexicaux fournis par sa langue, des moyens liés à la mise en forme matérielle du

texte. Celle-ci est donc à prendre tout à fait au sérieux comme composante à part entière du message. Le Modèle d'Architecture Textuelle, que nous présentons maintenant, offre un cadre théorique pour cela.

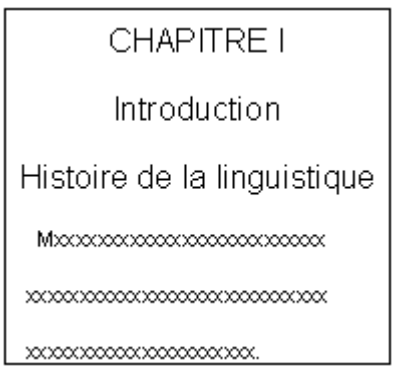
2.2. Prendre au sérieux la structure matérielle : le Modèle d'Architecture Textuelle

17 Parmi les fondements du Modèle d'Architecture Textuelle (par la suite MAT)[3], deux idées essentielles, illustrées par notre « jeu » de manipulations de la section précédente, sont à retenir : d'une part l'idée que les propriétés matérielles des textes traduisent les intentions de signification des auteurs du texte, d'autre part l'idée que tout ce qui est signifié par la mise en forme matérielle possède une relation d'équivalence avec une formulation discursive développée, c'est ce que nous avons mis en évidence dans notre commentaire de la figure 1. Ces deux aspects, que nous allons brièvement expliciter dans cette section, justifient de considérer les aspects visuels, c'est-à-dire les propriétés de mise en forme matérielle des textes, avec beaucoup d'attention et de les prendre en compte de la façon la plus fine possible dès lors que l'on vise des applications ayant pour but d'accéder au contenu des textes.

18 Pour le MAT, les éléments de mise en forme matérielle constituent des traces d'opérations réalisées par les *acteurs textuels :* « les acteurs textuels sont tous les intervenants dans la mise en forme matérielle du texte. Ils comprennent, entre autre : l'auteur, l'éditeur, le metteur en page, le typographe, etc. » . Ceux-ci, à leur choix, traduisent par une mise en forme spécifique leurs intentions de signification. Considérons, par exemple, une mise en forme matérielle telle que celle représentée par la figure 4.

19 Selon , une telle réalisation R1 peut être « mise en correspondance avec une phrase telle que la suivante : p1 : *J'introduirai ce premier chapitre par une histoire de la linguistique* » (*Ibid.*).

Figure 4. Image de page[4] adaptée de



20 Pour le MAT, la phrase p1 entretient avec la réalisation R1 une relation d'équivalence : elle « constitue une contrepartie discursivement développée de R1 » (*Ibid.*), c'est-à-dire qu'en lieu et place de R1 qui comporte titres et sous-titres, l'auteur du texte aurait pu se contenter de la phrase p1. Celle-ci est de type *métatextuel* : « les référents dont elle parle sont dans le texte même où elle apparaît et non des segments du monde dont parle le texte » (*Ibid.*). P1 est une instance du métadiscours, c'est-à-dire le discours dont les

référents sont dans le texte, et qui formule les opérations que l'auteur accomplit dans la construction de son texte.

21 Le MAT se fonde sur l'idée que « une des propriétés majeures du métadiscours associé à un énoncé est qu'**il peut être réduit dans une proportion aussi importante que permise par le contexte de la communication, tout en laissant des traces de cette réduction** [...]. On dira donc que R1 est le résultat de diverses réductions opérées sur la métaphrase p1. La nature des traces indique celle des réductions : transformations syntaxiques (« j'introduirai » / « introduction »), éléments lexicaux (« titre »), propriétés typo-dispositionnelles, mais aussi ponctuation. » (*Ibid.*, nous soulignons).

22 En résumé, le premier acteur textuel qu'est l'auteur organise son texte selon ses intentions de communication, par des opérations métatextuelles telles que celle explicitée par la métaphrase p1. Les métaphrases rendant compte des opérations d'organisation du texte peuvent être réduites à divers degrés (de 'aucune réduction' à 'réduction totale'), non sans laisser dans le texte des traces de diverses natures à partir desquelles le destinataire – le lecteur – construit son interprétation du texte.

23 La mise en forme matérielle, et notamment la typographie, la disposition, les effets tels que le soulignement, la mise en caractère gras, en italique, la numérotation, les puces, l'indentation, etc., ne sont donc pas dans ce modèle des éléments d'embellissement du texte, destinés à en améliorer la qualité esthétique, mais bien des éléments **fonctionnels**. C'est à ce titre qu'il est nécessaire de les intégrer dans les applications. Signalons d'ailleurs que, pour l'oralisation de l'écrit – par exemple pour offrir à des non-voyants une contrepartie orale de certains documents écrits –, F. Maurel analyse très soigneusement les caractéristiques morpho-dispositionnelles des textes à transposer automatiquement de l'oral à l'écrit, et propose un système apte à les prendre en compte.

24 Une précision s'impose toutefois : il n'y a aucune convention associée à la mise en forme matérielle des textes. Selon sa réalisation, selon le texte, une portion de texte en italique ou en gras n'aura pas la même fonction. Elle pourra, par exemple, signifier une insistance, ou introduire un nouveau concept ou un nouveau terme dans le texte, ou encore signaler un propos rapporté... On ne peut établir de correspondance de type bijectif entre une mise en forme matérielle et une intention de signification définie. L'important est le **contraste** entre différents éléments du texte[5] car ce contraste permet de définir les objets textuels : « un *objet textuel* est un segment de texte rendu perceptible par un jeu de contrastes de la mise en forme matérielle. Ainsi, parmi les objets textuels, on trouve les définitions, les énumérations, les parties, les titres, etc. » .

25 Certains des objets textuels mentionnés dans la citation qui précède, quoiqu'ils ne se signalent pas par une mise en forme matérielle invariable, sont particulièrement intéressants pour des tâches qui impliquent un accès au contenu du texte, comme l'acquisition de connaissances : les définitions permettent d'accéder aux connaissances du domaine, les énumérations permettent de repérer certaines informations, et nous verrons plus loin que les titres ont des fonctions diversifiées. La caractérisation de ces divers objets textuels pour des tâches de traitement automatique fait la part belle à leurs propriétés typo-dispositionnelles.

3. Exploiter la matérialité du texte pour le repérage d'informations spécifiques

26 De nombreuses tâches automatisées ayant pour objectif de fournir un accès au contenu du texte ne s'appuient encore que timidement, voire pas du tout, sur les propriétés matérielles des textes. Sont plutôt privilégiées des méthodes de statistique lexicale, ou la recherche de motifs linguistiques préalablement identifiés comme pertinents pour la tâche à accomplir, ou encore une combinaison des deux. Par exemple, associent une approche statistique et la recherche d'introducteurs de cadre thématiques afin de segmenter un texte en unités thématiques.

27 Il est vrai que les textes ne présentent pas toujours des indices matériels exploitables : nous avons vu dans la section précédente que les opérations de structuration peuvent tout à fait être exprimées par des formulations discursives, ce qui implique la recherche de marqueurs lexicaux et/ou syntaxiques. Mais nous avons aussi noté que ces marqueurs peuvent être réduits, et c'est précisément lorsqu'ils sont absents ou réduits que la matérialité du texte doit être pleinement exploitée.

28 Afin d'illustrer l'intérêt d'intégrer les aspects matériels des textes aux tâches de traitement automatique, nous prendrons comme exemple les recherches qui visent à acquérir des connaissances à partir des textes. Pour l'acquisition de connaissances, le repérage des définitions incluses dans les textes, formulées par les auteurs des textes eux-mêmes, est un enjeu important. Il est donc essentiel de modéliser finement les diverses formes sous lesquelles les définitions se présentent dans les textes, afin de construire des outils de repérage automatique.

29 Lorsqu'on se penche sur la description de leurs réalisations effectives en corpus[6], il apparaît que les définitions illustrent bien la gamme de possibilités qui vont d'une formulation discursive développée à une formulation visuelle. La figure 5 présente quatre exemples qui vont du « tout-discursif » au « tout-visuel ».

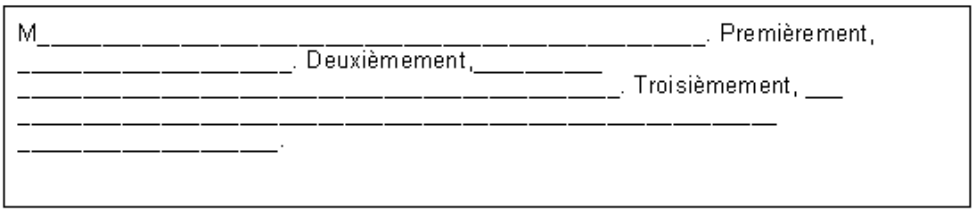
30 De la réalisation 1 à la réalisation 4 (qui ne couvrent pas toutes les possibilités, loin s'en faut), on voit se transformer puis disparaître les éléments lexico-syntaxiques qui permettent d'interpréter la définition comme telle : « je définis », « définition », et même, dans la dernière réalisation, la copule « est un ». Les éléments lexico-syntaxiques typiques de la définition sont suppléés ou remplacés par un marquage visuel de telle sorte que, dans R4, aucun des marqueurs linguistiques connus pour signaler une définition n'est présent. Seules des marques typo-dispositionnelles assurent le fonctionnement définitoire de l'énoncé : « la mise en valeur typographique du *définiendum* est pour nous un signal de son usage autonymique. La position en début de paragraphe, à la suite d'un titre [...], paraît également caractéristique » .

Figure 5. Différentes formulations d'une même définition

31 Il est donc indispensable, si l'on veut assurer le repérage des définitions dans les textes, de prendre en compte leurs caractéristiques typo-dispositionnelles au même titre que leurs caractéristiques lexico-syntaxiques : « [l'étude des définitions dans les textes] vise une modélisation fine des réalisations linguistiques des définitions, c'est-à-dire des réalisations de mise en forme matérielle grâce auxquelles ces objets textuels sont rendus perceptibles en discours. Les configurations de marqueurs qui doivent permettre de construire des filtres pour le repérage des définitions dans les textes incluent des marqueurs visuels : ponctuation, typographie, disposition. » .

32 Les définitions ne sont pas les seuls objets textuels dans la modélisation desquels le marquage visuel joue un rôle important. Pour les énumérations, la mise en forme matérielle se range parmi les indices essentiels. Même si, là encore, une énumération peut être marquée par des moyens strictement lexico-syntaxiques, c'est le plus souvent à des caractéristiques typo-dispositionnelles qu'elle se reconnaît. La figure 6 comporte deux exemples d'énumérations, la première est organisée par des marqueurs d'intégration linéaire (MIL) , la seconde par une typo-disposition spécifique renforcée par une numérotation.

Figure 6. Enumérations discursive et visuelle



33 De même que les définitions, les énumérations sont des objets textuels précieux dans la mesure où « énumérer c'est conférer une égalité d'importance à un ensemble d'objets » . Une énumération donne en effet accès à un ensemble d'items rassemblés sous un critère commun. Dans certains cas, elle permet d'atteindre une **collection** d'entités rassemblées sous une même étiquette : « L'identité de statut des constituants au sein de l'énumération exprime l'identité de statut des entités recensées dans le monde » .

34 Ces propriétés font donc des énumérations des objets textuels à privilégier lorsqu'on veut constituer certaines ressources lexicales à partir d'un corpus. C'est pourquoi en ont fait la cible privilégiée d'une recherche d'Entités Nommées (EN), « appellation générique pour les noms propres désignant des personnes, des lieux ou des organismes » . Pour la réalisation de cette tâche, leur approche a ceci d'original par rapport aux approches généralement utilisées en linguistique de corpus pour l'acquisition de connaissances qu'elle « exploite des informations sur la structure des documents pour acquérir des données » (*Ibid.*). Ceci est rendu possible par le fait que leur terrain d'investigation est constitué de pages web, donc au format html, et que ce format comporte des indications de mise en forme matérielle : les listes qui constituent les énumérations « visuelles », telles que le second exemple de la figure 6, sont soit intégrées dans des tables, soit marquées par des balises spécifiques (<*ul*> ou <*ol*>). La méthode d'acquisition intègre donc, parmi les analyseurs chargés de la moisson et du filtrage des pages web, un analyseur de liste qui s'appuie sur de telles marques pour repérer les EN candidates.

35 L'évaluation de la méthode (que nous n'exposerons pas plus en détail, se reporter à), met en évidence la nécessité de faire coopérer des analyses du lexique et des structures syntaxiques et une analyse des marques de structuration du document car nombre d'énumérations associent étroitement des marques lexicales et des marques visuelles.

36 Ces deux exemples empruntés au champ de l'acquisition de connaissances à partir de textes montrent de quelle façon la mise en forme matérielle des textes peut être mise à profit pour le repérage de segments textuels supposés receler un type spécifique d'informations. La recherche que nous présentons maintenant n'adopte pas tout à fait le même angle d'attaque. Il ne s'agit plus de modéliser les diverses formes sous lesquelles un objet textuel particulier se présente dans les textes afin de permettre son repérage, mais d'explorer la gamme de fonctions remplies par un objet textuel – les titres de section – afin de déterminer le type d'informations que l'on peut en tirer pour une caractérisation des contenus des documents.

4. Les titres de section : qu'indiquent-ils ?

37 En guise de préambule, on peut souligner l'importance accordée aux titres de section dans diverses tâches automatisées :

- pour le système d'aide à l'indexation IndDoc, le calcul de la pertinence d'un renvoi fait intervenir une pondération différente pour un terme apparaissant dans un titre ;
- pour la plate-forme ContextO, lors de la production de résumés automatiques de type Résumé-Auteur, « les titres de chaque section du texte original sont systématiquement placés dans le résumé, même si aucune phrase de la section n'a été sélectionnée » ;
- pour la résolution automatique d'anaphores basée par R. Mitkov sur le calcul d'une pondération des candidats antécédents, un poids supplémentaire est attribué au candidat situé dans un titre .

38 Les raisons de cette importance ne sont cependant pas toujours clairement explicitées, comme si le rôle des titres dans les documents allait de soi. Loin de nous l'idée de contester le bien-fondé de l'importance accordée aux titres de sections, au contraire, nous voulons l'asseoir sur une compréhension fine de leurs fonctions. Commençons par expliquer les motivations de notre recherche.

4.1. Motivations d'une recherche sur les titres de section

39 En tant que lecteurs, et particulièrement en tant que lecteurs de documents fortement structurés tels que des articles scientifiques, des comptes-rendus de recherche, des mémoires, des thèses, des projets, des ouvrages scientifiques, didactiques, *etc.*, nous savons, de façon plus ou moins intuitive, que les titres sont des éléments particuliers d'un texte. La table des matières d'un ouvrage volumineux n'est-elle pas le lieu par excellence en filigrane duquel s'inscrivent les thèmes abordés, les préoccupations et le point de vue de l'auteur ? Et la table des matières n'est autre chose que l'extraction et le regroupement de l'ensemble des titres du document (la *titraille*). Aux titres est associée une présupposition de pertinence : un lecteur s'attend à ce qu'ils soient appropriés à ce qu'ils titrent, c'est-à-dire à ce qu'il y ait entre le titre et l'objet titré une relation telle que le titre donne une information sur le contenu sémantique de l'objet titré. Généralement, les titres sont supposés être de bons indicateurs de la thématique du document.

40 Dans le cadre de traitements automatiques, un intérêt majeur de la compréhension du rôle des titres est lié à la (relative) facilité de repérage de ces objets textuels. En effet, dans un document numérique, les titres sont généralement identifiables grâce à un format spécifique. Par exemple, dans le traitement de texte Word, ils sont mis en forme avec un style de titre ; dans une page html, ils sont associés à des balises particulières (<h1> <h2>, etc.) ; dans un document xml, ils sont de même encadrés par des balises spécifiques dont le nom est défini par le concepteur du document (dans OpenOffice ou StarOffice, qui enregistre les textes au format xml, les titres sont balisés *heading 1*, *heading 2*, ...). Même dans des textes bruts, il est possible, avec certaines heuristiques, de repérer des titres de section[7] . Les titres sont des éléments formels des textes, et en tant que tels dotés de propriétés formelles assez aisément repérables. La possibilité d'un repérage automatique de ces objets textuels est donc une autre raison qui justifie de se pencher sérieusement sur l'analyse des fonctions des titres de section.

41 Notre recherche a pour objectif non seulement d'identifier les fonctions remplies par les titres, mais aussi d'inventorier les marqueurs formels de ces fonctions. L'hypothèse sous-jacente est qu'une caractérisation fine de chaque titre est susceptible d'enrichir la caractérisation des contenus d'un document. Nous envisageons « cet Objet Textuel spécifique qu'est le titre en termes d'instruction métadiscursive du scripteur à l'attention des lecteurs. » . Selon ce point de vue, les titres induisent certains processus interprétatifs et annoncent certains contenus. Nous pensons qu'il est possible de distinguer le type de contenu que le titre annonce par le prélèvement d'indicateurs dans son environnement.

4.2. L'étude des fonctions des titres : méthodologie

42 Nous avons basé notre étude[8] sur l'analyse d'un corpus de textes de trois origines différentes : des documents de travail produits dans le domaine de la gestion des déplacements, des articles scientifiques du domaine de l'ingénierie des connaissances (ces textes servent de substrat au projet CEDERILIC), des articles scientifiques du domaine de la géopolitique, téléchargés sur le site web de l'IFRI. Tous les textes sont sous format électronique et ont *a priori* été rédigés pour un format papier, ou pour un format mixte dans le cadre du projet CEDERILIC. Notre premier traitement de ces textes a consisté à baliser les titres, soit en nous appuyant sur les éléments de formatage conservés dans les documents numériques, soit en mettant en œuvre quelques heuristiques simples. Le corpus a été constitué de telle façon que chacun des trois sous-ensembles comporte un nombre comparable de titres : 345 / 348 / 348, ce qui donne un nombre total de 1041 titres.

43 L'ensemble des corpus a ensuite été intégré dans une base de données, les titres figurant dans une table spéciale. L'objectif à ce stade est d'enrichir chaque titre d'annotations concernant ses relations avec le reste du texte. Puisque nous voulons évaluer l'hypothèse que la caractérisation d'un titre peut permettre de caractériser le contenu du texte adjacent, il est nécessaire de cerner la façon dont le titre s'intègre – ou ne s'intègre pas – au texte. C'est pourquoi les annotations se sont focalisées sur le fait que les constituants d'un titre sont ou non repris dans la suite du texte. Les indications sur les reprises des titres nous paraissent éclairer le rôle joué par le titre. Par exemple, si un titre est repris par un pronom, cela témoigne à la fois d'une intégration syntaxique importante du titre et d'une capacité à introduire un référent dans le discours de telle manière que celui-ci est accessible à une anaphore pronominale, c'est-à-dire est placé au centre de l'attention, comme on le voit dans l'extrait suivant, où nous mettons en caractères gras la reprise pronominale du titre.

> [1] 2.4.3 Les données météorologiques
> **Elles** sont recueillies par des stations météoroutières.

44 Si on compare cet exemple avec l'extrait reproduit en [2], dans lequel aucun des mots présents dans le titre ne réapparaît dans la section titrée, on voit nettement se dessiner des modalités fondamentalement différentes d'intégration du titre dans le texte, qui distinguent des fonctions hétérogènes du titre, nous les détaillerons dans la section suivante.

> [2] b. Ascension et chute de John D. Rockefeller
> De très nombreux entrepreneurs, souvent des aventuriers risquant leur fortune personnelle, tentèrent leur chance dans l'exploration pétrolière à partir des années 1860. Acquérant des droits sur de minuscules parcelles ou sur des milliers d'hectares, ils furent les acteurs de l'ère héroïque de l'histoire pétrolière

> américaine. Mais l'amont pétrolier (exploration et production) est une activité extrêmement risquée et pour beaucoup de ces pionniers l'expérience tourna court. La plupart ne découvrirent rien mais ceux qui eurent la chance d'accéder au stade de la production affrontèrent la dure réalité d'un marché libre de matière première. Le développement intensif des premières découvertes précipita rapidement une chute du prix du pétrole, qui passa de $ 37 à $ 7 entre 1870 et 1890. S'ensuivit une vague de faillites et un mouvement de consolidation de l'industrie (sélection et concentration).

45 Pour clore sur notre méthodologie, précisons les éléments d'annotation des titres. Pour chaque titre du corpus, nous avons noté manuellement :

s'il fait ou non l'objet d'une reprise ;
si oui, on précise la forme de la reprise :

- elle est totalement identique, c'est-à-dire que le titre est répété dans la section titrée sans la moindre modification ;
- elle concerne uniquement les mots lexicaux[9] : ceux-ci réapparaissent dans le texte, mais avec une modification (par exemple, changement de nombre) ou/et dans un ordre différent de celui du titre ;
- elle concerne seulement une partie des mots lexicaux ;
- elle est pronominale ;
- elle est une construction présentative ;

– s'il y a reprise, on précise la localisation de la reprise :

- elle se situe dans la première phrase du paragraphe qui suit le titre ;
- elle se situe dans la suite du paragraphe ;

on indique si la reprise est le *topic*[10] de la phrase dans laquelle elle apparaît, ce que nous avons considéré en prenant en compte essentiellement la position syntaxique de sujet ;
on note s'il y a ou non un autre titre qui suit immédiatement le titre analysé ;
s'il y a un autre titre immédiatement après, on note si on a ou non une reprise du titre analysé dans cet autre titre.

46 Ce dernier point nous permet de capter les successions de titres dont chacun reprend et précise ce qui est mentionné dans le titre qui précède :

> [3] Les mesures anti-terroristes aux Etats-Uni depuis septembre 2001.
> 1/ Les premières mesures dans le sillage des attentats.
> Avec le vote du USA Patriot Act (PL 107-56) dès octobre 2001, Bush et ses conseillers ont considérablement renforcé les textes déjà existant concernant la lutte contre le terrorisme [...].

47 Dans l'extrait ci-dessus, *mesures* est repris dans le titre de niveau 2, ce dernier opérant une restriction sur l'ensemble des *mesures anti-terroristes aux Etats-Unis depuis septembre 2001*.

48 Les annotations effectuées manuellement sur les 1041 titres du corpus d'étude nous permettent à l'heure actuelle de construire une ébauche de typologie des titres.

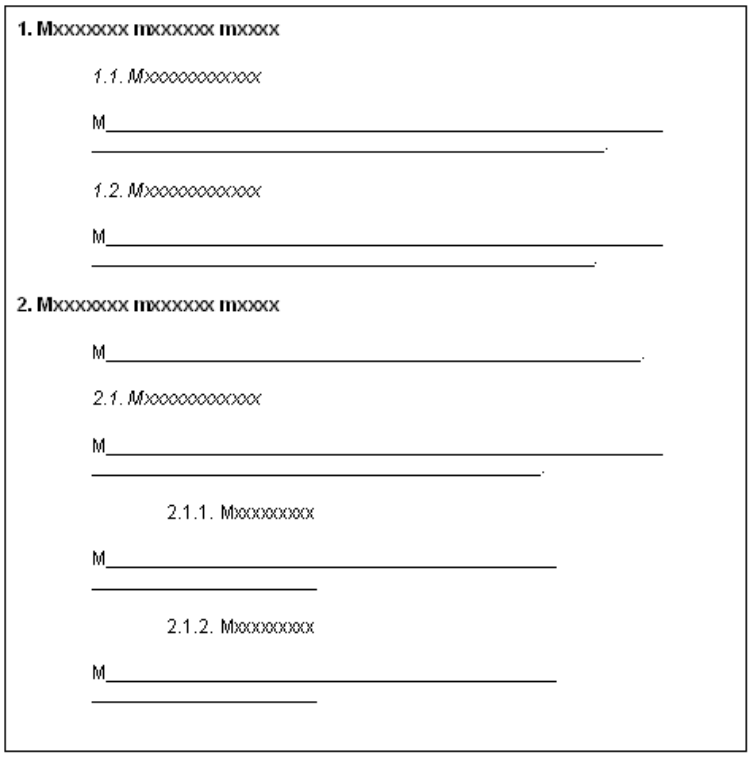
4.3. L'étude des fonctions des titres : résultats

49 On peut globalement dire que les titres de section dans des documents textuels jouent un rôle sur deux plans, sur le plan matériel et sur le plan notionnel.

50 **1. sur le plan matériel**, ils subdivisent et en même temps regroupent et hiérarchisent. Avant même de prendre en compte leur contenu sémantique, les titres organisent le texte d'une façon immédiatement **visible**, ce que montre l'image de page figure 7.

51 Le type de segmentation illustré par la figure 7 joue un rôle bien au-delà d'un repérage visuel de la structure du texte. En donnant à voir une hiérarchisation des sections, c'est déjà une structuration d'ordre sémantique qui est construite : les places qu'occupent les différentes sections dans la structure globale du texte sont à traiter en termes de relations de subordination, de juxtaposition, de coordination, qui sont plus précisément spécifiées par le contenu notionnel des titres. Le lecteur est sensible à l'emboîtement et au parallélisme qui lui permettent de construire mentalement le squelette du discours.

Figure 7. Organisation visuelle du texte par les titres de section



52 En même temps que la structuration/hiérarchisation, les titres procurent une dénomination pour les sections qu'ils intitulent. On peut ainsi faire référence à telle section du document en la désignant par le titre qui lui a été donné dans le document, ce qu'illustre l'extrait suivant :

> [4] 2. LA COOPERATION SUR L'AGGLOMERATION TOULOUSAINE
> [...]
> 3. LE SYSTEME DE GESTION GLOBALE DES DEPLACEMENTS (SGGD)
> Le SGGD a été présenté au chapitre A (**cf. le paragraphe « La coopération sur l'agglomération toulousaine »**). Rappelons qu'il s'agit d'une démarche [...]

53 Le repérage de tels renvois, qu'ils s'effectuent au moyen de la répétition du titre comme ci-dessus, ou par référence à la numérotation interne du document lorsqu'elle existe – par exemple, *cf. §2.* –, permet de mettre en relation des segments disjoints au sein d'un document. Ces renvois s'apparentent à un lien hypertextuel : ils incitent le lecteur à s'affranchir de la linéarité du document papier.

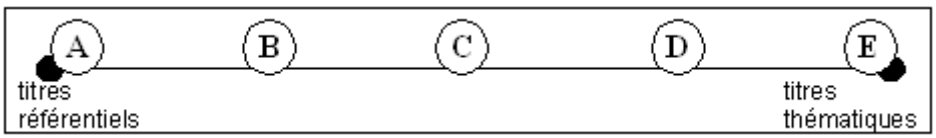
54 **2. sur le plan notionnel**, c'est-à-dire en ce qui concerne la construction du **contenu** du texte, les titres assument une palette de fonctions qui vont de l'introduction (ou la préparation à l'introduction) de nouveaux référents dans le discours à l'introduction

d'une thématique, en passant par la focalisation sur certains référents. Ce sont ces fonctions que nous exposons maintenant (section 4.3.1), puis nous indiquerons en quoi la caractérisation de types de titres peut aider à la caractérisation du contenu du document (section 4.3.2).

4.3.1. Une typologie des titres selon leur rôle

55 Selon l'intuition ordinaire, les titres de section se subdivisent en deux grandes catégories : certains annoncent des sections typiques d'un certain genre de texte (scientifique, didactique), ce sont les *Introduction*, *Conclusion*, *etc.* ; les autres peuvent être vus comme ayant pour fonction essentielle d'indiquer la thématique[11] de la section qu'ils intitulent. Lorsqu'on examine un corpus tel que celui que nous avons constitué, il apparaît effectivement que cette fonction « thématique » est extrêmement présente. Mais elle n'est pas le seul rôle possible du titre. Selon notre analyse, cinq catégories se dessinent, réparties sur un continuum dont les pôles extrêmes sont constitués de ce que nous appellerons *titres thématiques* et de ce que nous appellerons *titres référentiels* (pour cet article, nous laissons de côté les titres du type *Introduction*, *Conclusion*, ils sont pris en compte dans).

Figure 8. Typologie des titres en cinq catégories



56 A un pôle du continuum, les titres référentiels sont un ensemble de titres dont la fonction essentielle est de préciser le ou les référents dont la suite du texte va parler. A l'autre pôle, les titres thématiques n'introduisent pas dans le discours un référent précis, mais plutôt délimitent un cadre thématique dans lequel s'inscrit ce dont on va parler : un domaine d'activité, un domaine de connaissances, un point de vue, une situation spatio-temporelle, etc., spécifiques. Ces deux pôles renvoient à des processus interprétatifs différents : il s'agit dans le premier cas, d'attirer l'attention du lecteur sur un ou des référents du discours particulier(s), dans le second, de canaliser certaines de ses connaissances d'arrière-plan. Entre les deux pôles, les catégories (B), (C), (D) constituent un mixte de ces deux extrêmes. Nous caractérisons chacune des catégories par la fonction qui y domine :

- des titres qui essentiellement jouent un rôle dans la gestion des référents du discours ;
- des titres qui jouent un rôle dans la gestion des référents du discours, tout en jouant un rôle dans la définition de la thématique textuelle ;
- des titres mi-référentiels, mi-thématiques ;
- des titres qui jouent un rôle dans la définition de la thématique textuelle, tout en ayant aussi un rôle dans la gestion des référents ;
- des titres qui essentiellement jouent un rôle dans la définition de la thématique textuelle.
- Revenons plus précisément sur chacun de ces types et sur les indices qui permettent de les distinguer.

57 **Le type (A)** représente les titres référentiels par excellence, c'est-à-dire les titres exclusivement voués à la gestion des référents du discours. Une caractéristique formelle essentielle de ces titres, liée à leur fonction, est qu'ils sont formés de syntagmes

nominaux, parfois réduits à un nom seul. Ils interviennent dans la gestion des référents selon trois modalités.

1. Le titre focalise l'attention du lecteur sur un référent déjà installé dans le discours.

58 Dans l'extrait qui suit, le *plan de directive* a été introduit, sous une forme quelque peu différente, dans le titre *7.3.1*. Le titre *7.3.1.2* le remet au premier plan, au centre de l'attention. Le référent peut ensuite être repris comme *topic* de la première phrase de la section.

> [5]7.3.1 PLAN DE REFERENCE D'UNE DIRECTIVE
> 7.3.1.1 PAGE DE GARDE OU PAGE DE COUVERTURE
> [...]
> 7.3.1.2 PLAN D'UNE DIRECTIVE
> **Ce plan de directive** est communiqué à titre de référence, c'est-à-dire qu'**il** indique le niveau d'information à donner. [...]

59 2. Le titre introduit un nouveau référent dans le discours.

60 Dans ce cas de figure, le référent n'est pas mentionné dans le texte antérieur.

> [6] Le programme de défense anti-missile
> **Le programme de défense anti-missile** (MD) reçoit 7.5 milliards de dollars dans le projet de budget pour 2003.

61 La distinction entre ces deux premiers types repose exclusivement sur le repérage d'une introduction préalable du référent exprimé par le titre, car pour ce qui est des indices formels liés aux reprises, ils sont similaires. En effet, la reprise intervient immédiatement après le titre, elle concerne soit la totalité du titre, soit le nom tête du syntagme nominal qui constitue le titre (par exemple, *Bulletin prévisionnel* est repris par *Ce bulletin*). Elle est généralement le sujet de la phrase et adopte un déterminant défini ou démonstratif.

62 3. Le titre prépare l'introduction d'un référent dans le discours.

63 Le référent exprimé dans le titre fait l'objet d'une introduction, en position saillante, dans la première (ou éventuellement la seconde) phrase du paragraphe. Ce n'est qu'après cette introduction, liée souvent à une explicitation ou une justification de ce que le référent a à voir avec le propos global, que ce référent devient le *topic* des phrases qui suivent.

> [7] 5.3. La réutilisation L'une des techniques proposées pour faciliter le processus de modélisation, en ingénierie des besoins comme en ingénierie des connaissances, est **la réutilisation** de modèles. **Elle** devient un objectif prépondérant. **Il s'agit de** réutiliser des modèles (ou des parties de modèles) conçus sous une forme générique, précédemment développés et stockés dans des bibliothèques spécialisées.

64 On voit dans l'extrait ci-dessus qu'après son introduction dans la première phrase de la section, *la réutilisation (de modèles)* est repris par un pronom, puis fait l'objet d'un énoncé définitoire marqué par *il s'agit de...*

65 Ces titres préparatoires se reconnaissent à la répétition du titre dans une position autre que sujet de la phrase, avec une variation limitée à un éventuel changement de déterminant (par exemple, article indéfini / article défini).

66 **Le type (B)** reste près du pôle référentiel, selon les modalités que nous venons d'exposer. Mais tout en assurant la gestion d'un référent, ces titres commencent aussi à dire quelque chose de ce référent : ils en prédiquent quelque chose, ou indiquent le point de vue selon lequel on va en parler, etc. Ces titres aussi sont essentiellement des syntagmes nominaux.

> [8] C -- Les agences fédérales civiles, utilisatrices concurrentes
> Au-delà de la communauté du renseignement et des forces armées américaines, **les**

agences fédérales civiles font également usage d'imagerie spatiale pour mener à bien leur mission.

67 Par rapport au type référentiel « pur », ces titres présentent la particularité de pouvoir être segmentés en deux parties : une première partie exprime le référent, une seconde partie exprime une prédication ou un cadrage thématique par rapport à ce référent, et c'est la première partie qui fait l'objet d'une reprise.

68 **Le type (C)** associe « à parts égales » une fonction référentielle et une fonction thématique.

[9] 3.4.3 COMPARAISON DES SOLUTIONS
Les différentes **solutions** envisagées seront **comparées** selon les critères suivants :
* coût global (études, rréalisation, maintenance),
* évolutivité (en précisant les axes d' évolution), [...]

69 Le titre introduit un référent qui est repris dans la première phrase, les *solutions*, mais sous l'angle de la *comparaison*, repris sous la forme verbale *comparées*.

70 La différence de ces titres avec les titres (B) réside dans le fait que des reprises des différents mots du titre se répartissent au fil du texte, avec, comme on le voit ci-dessus, une possibilité de variation morphologique de certains mots du titre. Cette répartition des reprises dans le texte contribue à constituer la thématique de la section.

71 Avec l'exemple illustrant **le type (D)**, qui classe des titres dont la fonction première est la délimitation d'un cadre thématique mais qui jouent tout de même un rôle dans la gestion des référents, nous pouvons voir que la répartition des reprises est un des traits marquants des titres qui se rangent du côté du pôle thématique. Ils font volontiers l'objet de reprises éparses des mots qui les constituent, quoique ce ne soit pas une condition nécessaire.

72 Un autre trait est lié à leur forme, ils regroupent des formes variées telles que syntagme prépositionnel, syntagme verbal, ou encore phrase. Dans les cas, toujours possibles, où ce sont des syntagmes nominaux, ceux-ci sont alors des syntagmes nominaux bipartites, leur bipartition étant marquée par une coordination ou par une ponctuation interne (virgule, point-virgule, deux-points, tiret, etc.), comme le montre l'extrait qui suit.

[10] 4.2.3. Une première formalisation des prototypes : pondération des attributs et valeurs
Nous avons dégagé, à partir de ces analyses des similarités extensionnelles, quatre **prototypes** correspondant globalement à des comportements différents. Ce sont ensuite les fréquences de cooccurrence des valeurs d'attributs qui nous ont permis **de pondérer attributs et valeurs** pour chacun de ces comportements.

73 Le titre prépare l'introduction du référent de *prototypes*, qui n'est réellement installé dans le discours que par la première phrase du paragraphe. Cependant, la fonction première de ce titre n'est pas référentielle, elle est de canaliser l'attention du lecteur sur la *formalisation des prototypes* et plus particulièrement sur ce qui constitue la *première formalisation*, c'est-à-dire la *pondération des attributs et des valeurs*. Le titre ouvre un espace thématique qui est ensuite déployé dans la section, ce que montrent les diverses reprises des mots constituant le titre. D'une certaine manière, de tels titres condensent le contenu de la section titrée pour délimiter, canaliser les connaissances et inférences qui devront être mobilisées par le lecteur pour une interprétation de ce qui suit.

74 **Le type (E)** représente le parangon du titre thématique.

[11] 4.1.3. D'un point de vue technique
Premièrement, l'interface a été conçue pour inciter les étudiants à utiliser certains

> outils (même si l'on savait à l'avance que, en toute hypothèse, les étudiants font ce qu'ils veulent ; ainsi, certains ont utilisé un outil externe de mail) ; en l'occurrence, parce que nous étions focalisés sur la dynamique du groupe, nous les avons incités à utiliser des outils de travail synchrones lors de phases collectives (aspects « naturel et convivial »), alors que le synchrone n'est généralement pas synonyme d'accélération d'une prise de décision commune. Deuxièmement, l'articulation des outils synchrones et asynchrones ne fait pas l'objet d'un dispositif technique, mais de l'intervention d'un des étudiants (le rédacteur), rôle attribué par émergence. Enfin, la circulation des données entre les étapes ne fait pas l'objet d'un dispositif technique ; elle est gérée par le tuteur, afin que celui -ci soit partie intégrante de l'activité.

75 Ce titre n'introduit aucun référent dans le discours, il indique au lecteur que ce qui suit relève d'un certain point de vue (en fait, c'est ce qui est introduit dans le titre 4.1., *Spécificités du contexte pédagogique*, qui est, tout au long de la section, examiné selon divers points de vue) et, ce faisant, lui fournit un cadre interprétatif pour l'ensemble du propos subséquent.

76 L'ébauche de typologie que nous venons de présenter dans cette section ne prétend pas épuiser le sujet, elle vise à couvrir les types de titres présents dans notre corpus. Il reste à vérifier dans quelle mesure elle peut se montrer appropriée aux titres de journaux, aux titres de chapitres de romans, aux titres de poèmes ou à d'autres sortes de titres... Et pour les documents de type informatif tels que ceux qui constituent notre corpus, sans doute serait-elle enrichie et/ou affinée par l'intégration de documents relevant de registres sensiblement différents. Elle nous permet toutefois d'indiquer comment l'identification d'un type pour un titre donné est susceptible d'enrichir l'accès au contenu d'un document.

4.3.2. Accéder au contenu par le type de titre

77 Le parti que l'on peut tirer d'une telle typologie réside dans le fait que le type d'un titre constitue un indicateur au moins partiel du contenu de la section titrée, ce qui permet de cibler des zones potentiellement propices à l'expression de certaines informations.

78 Les titres à composante référentielle, et particulièrement les titres qui assurent l'introduction d'un nouveau référent dans le discours, peuvent être liés à une définition :

> [12] 3.3. Normalisation
> **La normalisation** est un processus particulier de conceptualisation fondé sur l'analyse de corpus [...]. **La normalisation** consiste en une interprétation sémantique dans le but de structurer des concepts et des relations sémantiques.
> [13] 2.7 Les systèmes embarqués d'information des automobilistes
> **Ces systèmes embarqués** se regroupent en deux catégories principales, ceux qui reçoivent et émettent (notamment les radiotéléphones et les postes de CB), ceux qui reçoivent seulement (postes radiophoniques, balises, informatique embarquée de guidage).

79 Même si le début de la section ne comporte pas une définition « canonique », une section ouverte par un titre référentiel livre généralement des informations à propos du référent mentionné dans le titre :

> [14] 2.4.2 La visualisation du trafic
> **Cette visualisation** s'effectue essentiellement par le suivi vidéo en temps réel.

80 De même, ces titres à composante référentielle peuvent former le premier maillon d'une chaîne topicale[12] :

[15] 2.4.5 Le réseau d'appel d'urgence
Ce réseau, géré par la DDE et ASF, est un pourvoyeur direct de renseignement sur les incidents. **Il** est alimenté par les communications d'usagers à partir des postes d'appel d'urgence.

81 Ces remarques s'étendent, dans une moindre mesure, à des titres qui ne sont pas « purement » référentiels, c'est-à-dire qui ne se classent pas comme les précédents dans la catégorie (A) mais peuvent se ranger en (B) ou même en (C). Les deux extraits qui suivent montrent comment, après un titre à la fois thématique et référentiel,

des éléments d'information sont apportés sur le référent :

[16] « Interim Force » : les brigades interarmes intermédiaires[13]
La brigade interarmes intermédiaire, dite IBCT (« Interim Brigade Combat Team »), a été lancée en octobre 1999 et vise deux grands objectifs : préparer la voie aux systèmes et aux formations futures de l'« Objective Force », et corriger les déficiences constatées récemment en matière de déploiement rapide et de « versatilité » des forces terrestres américaines, ce qui implique une réorganisation des structures.

une section entière peut être liée par une chaîne topicale (en gras) dont le premier maillon est introduit dans le titre :

[17] Fonctions du comité de pilotage
Il approuve les documents, [...], qui lui sont soumis par l'équipe de conduite d'opération, ainsi que les cahiers des charges [...]. Sur la proposition de l'équipe de conduite d'opération, **il** constate la mise en service du système ERATO. **Il** définit les modalités d'application de la politique de gestion du trafic automobile [...]. **Il** valide les principes de coordination entre les différents partenaires.

82 De telles chaînes délimitent des segments de textes homogènes sur le plan thématique et peuvent donc constituer un appui pour une segmentation automatique du texte.

83 La caractérisation de titres comme référentiels est ainsi susceptible d'apporter un plus dans une perspective d'acquisition de connaissances, ou tout simplement d'identification de zones de textes favorables à la récolte d'informations à propos des référents du discours.

84 Dans l'objectif d'une visualisation dynamique et interactive des textes (que nous avons évoquée dans l'introduction), ce type de titre peut apparaître ou disparaître de la représentation du texte selon le projet de lecture. Si le texte est envisagé selon une granularité forte (par exemple, la représentation des thèmes principaux), l'inclusion de ces titres dans la visualisation n'est guère pertinente. Si au contraire le texte est envisagé selon une granularité fine (les objets et référents du discours), alors ces titres référentiels trouvent toute leur place dans la visualisation.

85 Les titres dits thématiques seraient, eux, particulièrement pertinents pour visualiser les idées et thèmes principaux du texte. Leur fonction est d'ouvrir un univers de discours ou de construire un champ thématique constituant un cadre pour ce qui va suivre. Ils entraînent donc dans leur sillage les différents éléments (référents, processus, etc.) composant l'univers de discours.

86 Par exemple, dans l'extrait suivant, le titre « annonce » que la section va préciser tous les ingrédients nécessaires à une mondialisation durable. On voit en effet dès la fin du premier paragraphe et dans le deuxième (nous n'avons pas reproduit la section dans son intégralité) surgir ces ingrédients : *des principes et des institutions de gouvernance globale*, *les gouvernements nationaux*, *accroître la transparence*, *accroître la possibilité pour certains*

représentants de la société civile d'exprimer leurs préoccupations, la réflexion sur les politiques nationales, etc.

> [18] Pour une « mondialisation durable »
> La mondialisation est un processus hétérogène, inégal selon les secteurs, et dont certains pays restent largement exclus. Il peut être mis au service de la croissance et du développement économique, à condition d'être encadré par des principes et des institutions de gouvernance globale, et aussi d'être promu par les gouvernements nationaux. Les réactions contre la mondialisation à la fin des années
> 1990 ont souligné le caractère incomplet des institutions de gouvernance globale (notamment en matière d'environnement), et le manque de transparence et d'ouverture des institutions économiques internationales. L'une des voies d'évolution consiste à accroître la transparence de ces institutions et la possibilité pour certains représentants de la société civile d'exprimer leurs préoccupations au cours des processus de décision. Cette évolution est désormais amorcée et doit être poursuivie. Au-delà, la gouvernance globale suppose à la fois de nouvelles institutions et une meilleure articulation entre certaines institutions existantes. Mais les politiques nationales sont tout aussi fondamentales pour soutenir le mouvement d'ouverture et promouvoir ses conséquences positives. La réflexion sur la gouvernance globale risquerait d'être une fuite en avant si elle se substituait à la réflexion sur les politiques nationales.

87 On a donc avec un titre tel que celui-ci un point de vue particulier à partir duquel saisir l'ensemble de ce qui est dit dans la section, et dont il faut tenir compte pour la représentation du contenu du texte ; dans cet exemple, alors que l'expression *politiques nationales* est présente dans divers endroits du document dont l'extrait est issu, elle acquiert dans le cadre instauré par le titre une valeur nouvelle.

88 A notre sens, la caractérisation des titres vient en complément d'autres techniques d'analyse textuelle. Elle ne fournit pas à elle seule les moyens d'un accès au contenu du document, mais permet d'affiner la représentation de celui-ci. Ce que nous avons esquissé jusqu'ici montre surtout que tous les titres dans les documents informatifs n'ont pas les mêmes fonctions et que leur inclusion dans une représentation partielle du contenu du texte doit être modulée selon ces fonctions. Nous en déduisons qu'il n'est sans doute pas pertinent de mentionner systématiquement la totalité des titres d'un document quand on veut en construire une représentation, et que les titres, comme le reste du texte, peuvent et doivent être filtrés.

89 Comme nous venons de le montrer, les titres jouent un rôle d'organisateurs textuels, à la fois en termes de découpage de la matière du texte et de structuration de son contenu. Afin de saisir plus complètement leur contribution à la construction du sens du texte, ainsi que celle d'autres éléments de structuration matérielle du texte, et dans l'objectif d'inclure davantage les informations de structuration visibles à la surface du texte dans un traitement automatique visant un accès au contenu textuel, il est essentiel d'une part de mieux cerner les fonctions de ces divers éléments, d'autre part de déterminer quels types d'informations de mise en forme matérielle doivent enrichir le texte sous format numérique.

5. Structure matérielle / structure sémantico-logique : des relations à explorer

90 Même s'il existe au moins un modèle tel que le MAT (cf. section 2.2) pour rendre compte du fait que les propriétés de mise en forme matérielle des textes participent à leur signification, même si, intuitivement, tout lecteur compétent ressent que le découpage en sections et en paragraphes, les alinéas, retraits et autres formes de segmentation remplissent véritablement une fonction de guidage de l'interprétation , il n'existe à l'heure actuelle aucune description opératoire élucidant les fonctions de ce découpage. A l'heure où le document écrit devient numérique, c'est-à-dire se présente sous des formats incluant diverses informations de structuration tels que xml ou html, il devient urgent de savoir quoi faire de ce type d'informations. Nous suggérons ici quelques pistes de recherche.

5.1. Sections « sœurs », quelles relations ?

91 Par exemple, s'agissant des titres et du découpage en sections qu'ils opèrent, la compréhension des différents types de relations qui s'instaurent entre les sections du texte permettrait de rendre compte d'une façon plus précise de la structuration du texte. Des sections d'un même niveau au sein d'une section supérieure peuvent en effet entretenir diverses relations. Sans viser l'exhaustivité, évoquons-en quelques-unes à titre d'exemple.

1. Relation de parallélisme, les titres tendent alors à ressembler aux items d'une énumération, comme dans l'extrait suivant où nous ne reproduisons que les titres et non le texte :

[19] Préparer la réponse en cas de diffusion d'imagerie métrique
a - Les campagnes militaires
b - La sécurité du territoire
c - Les médias et les ONG

92 Les trois titres numérotés *a*, *b* et *c* mentionnent les trois points sur lesquels *préparer la réponse*... Dans un tel cas de figure, aucune de ces trois sections n'a prééminence sur l'autre, on peut les considérer comme occupant des espaces similaires dans le texte.

2. Relation d'anaphore, un des titres de la section renvoie anaphoriquement à l'un des titres précédents :

[20] 1.2 La description du réseau actuel et ses évolutions
1.2.1 Les évolutions à court terme
1.2.2 D'autres évolutions notables à moyen terme

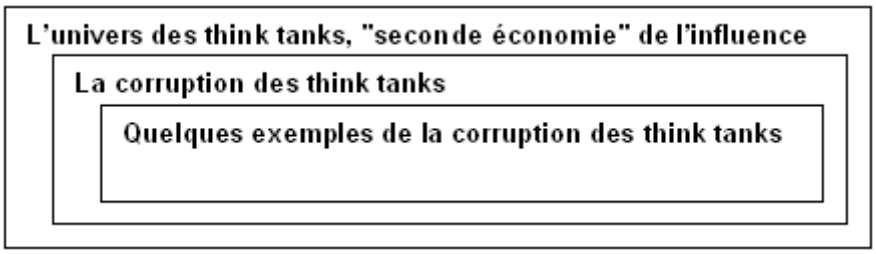
93 Le titre numéroté *1.2.2*, comportant l'adjectif *autres*, est nécessairement le deuxième titre au moins d'une section : il a besoin d'un précédent auquel « raccrocher » cet adjectif *autres*. De ce fait, le contenu de la section est nécessairement second, ce qui ne signifie pas secondaire ou subalterne, mais signifie que ce contenu entretient une relation de contraste ou de complément à l'égard d'un contenu forcément déjà présent. Contrairement à l'exemple précédent, les sections n'occupent pas des espaces similaires, celle dont le titre comprend *autres* présuppose qu'elle vient après une ou plusieurs sections de même niveau.

3. Relation d'emboîtement, chacun des titres reprend une partie du précédent (de même niveau) en le restreignant :

[21] L'univers des think tanks, "seconde économie" de l'influence politique
La corruption des think tanks
Quelques exemples de la corruption des think tanks

94 Quoique ces trois titres soient tous de même niveau, ils réalisent des découpages successifs au sein de l'univers instauré par le titre précédent. Nous pouvons représenter ceci sous forme d'emboîtements :

Figure 9. Relation d'emboîtement entre sections de même niveau



95 On peut noter ici que ce sont des titres de niveau 1, ce qui n'est sans doute pas étranger à cette configuration. En effet, leur intégrant n'est autre que le document lui-même, on voit que ces titres retracent le cheminement d'un propos qui se construit et se précise.

4. Relation de « conclusion », le dernier titre rassemble les contenus précédents selon une perspective particulière

[22] 3.2. Continuité sémantique
3.2.1. Sémantique interprétative
3.2.2. Sémantique opérationnelle
3.2.3. Sémantique formelle
3.2.4. Articulation des différentes sémantiques

96 Là encore, le dernier titre ne peut être à une autre place que celle de dernier titre. Le contenu de la section qu'il titre fait référence à et s'appuie sur le contenu des trois sections précédentes. Si celles-ci semblent occuper des espaces similaires, cette dernière section, se nourrissant de ce qui la précède, englobe en quelque sorte les trois premières.

97 Ces quatre exemples, nullement exhaustifs, montrent que le fait pour certaines sections d'être « sœurs » au sein d'une section de rang plus élevé n'implique pas un seul et même type de relation entre elles. Il se dessine ici un parallèle avec les énumérations qui recouvrent des modes de structurations variées . Si les sections et sous-sections constituent un découpage du texte en blocs perceptibles dans la matérialité du texte, la façon dont ces blocs se succèdent, se précisent, se reprennent, se modifient, etc., au sein d'un bloc plus englobant qu'est la section supérieure n'est pas uniforme.

98 Si on veut considérer un document non de façon plate, mais comme un objet structuré, et si on veut exploiter cette structure pour un accès sélectif au contenu du document, l'exploration des fonctions des éléments de structuration des textes tels que les titres, les sections, les paragraphes, ..., reste à faire. Au moins deux études récentes allant dans cette direction sont à signaler.

5.2. Deux études des interactions entre marques de structuration

99 M. Charolles , notant que « les travaux d'analyse linguistique du discours ne réservent en général aucune place à la mise en forme graphique et visuelle des textes. », envisage les paragraphes[14] comme un type particulier de marques de cohésion discursive. Il les analyse comme « des cadres organisationnels sous-spécifiés sémantiquement (quand ils n'ont pas de « titre ») » et s'intéresse particulièrement à leurs interactions avec d'autres marques de cohésion comme les anaphores, les marqueurs d'intégration linéaire (que nous avons évoqués dans la section 3) ou encore les cadres énonciatifs (par exemple *selon X*, *d'après Y*...), pour n'en citer que quelques-unes.

100 Dans le même esprit d'analyse de niveaux différents de la structuration du texte et de leurs interactions, M. Laignelet s'attache à « la structuration temporelle des discours à travers deux objets textuels particuliers, les titres et les adverbiaux à l'initiale de la proposition (ou introducteurs de cadres) » en se demandant quelles relations fonctionnelles entretiennent ces deux objets textuels.

101 Elle montre notamment qu'il existe un rapport entre le caractère temporel d'un titre et la présence et le type d'introducteurs de cadre temporels qui segmentent le texte sur l'axe temporel. Elle note – ce qui corrobore certaines de nos observations sur les titres – que « le titre [temporel] fonctionne effectivement comme une annonce (thématique au sens large) » (*Ibid.* : 164) et que, dans le segment titré, les introducteurs de cadre précisent, modulent ou découpent la période temporelle initiée par le titre.

102 Dans la perspective d'une navigation à l'intérieur d'un document, de telles études qui s'attachent à mettre en rapport des marques différentes de structuration, repérables en mettant en œuvre des moyens complémentaires, nous paraissent ouvrir de réelles voies d'accès aux contenus d'un document.

103 Mais il ne s'agit pas seulement à notre sens de dresser un inventaire des fonctionnements possibles, il faut en outre mesurer de façon précise comment ceux-ci sont (ou ne sont pas) déterminés par le genre textuel.

5.3. Fonctions des éléments de structuration et registre de texte

104 Pour donner un exemple de l'influence du genre, revenons à notre corpus d'étude (décrit dans la section 4.2). Dans celui-ci, chacun des trois sous-ensembles ne comporte pas les mêmes quantités des différents types de titres identifiés (décrits dans la section 4.3.1). On peut faire l'hypothèse, intuitivement plausible, que la fréquence d'emploi de ces différents types est corrélée à ce que D. Biber appelle le « registre de texte », c'est-à-dire que les textes privilégient tel ou tel type de titre en fonction de leurs caractéristiques fonctionnelles et sociales : le public visé, les intentions de l'auteur, la fonction du document (convaincre, expliquer, décrire, informer, former, ...), etc.

105 Malgré notre quantité limitée de données, nous voyons déjà se dessiner des différences remarquables, notamment dans l'emploi de formes de titres ou de configurations « marginales ».

106 Premier exemple, certaines formes de titres caractéristiques d'une fonction thématique, comme par exemple les titres qui ont la forme d'une phrase, sont significativement sous-représentées dans les textes professionnels, voire même pas représentées du tout. Nous

avons dans notre corpus en tout et pour tout 35 titres de forme phrase, ce qui justifie que nous la qualifiions de marginale. C'est essentiellement dans les textes du domaine géopolitique qu'ils se concentrent : 63 % des occurrences ; les autres occurrences se trouvent **toutes** dans les articles scientifiques du domaine Ingénierie des connaissances. Un test de Chi-deux, significatif à une probabilité <.001, montre que cette répartition n'est pas due au hasard. Nous l'interprétons comme un premier indice d'une corrélation entre forme / fonction du titre et registre de texte.

107 Deuxième exemple, la configuration « titre + reprise anaphorique par pronom » suit une distribution totalement inverse. Marginale elle aussi, puisque notre corpus n'en fournit que 6 occurrences[15], elle se trouve massivement (84 %) dans les textes professionnels.

108 Le fait qu'on soit d'un côté, face à des textes essentiellement informatifs, des textes de travail qui décrivent un état de choses (principalement comment s'organise la gestion des déplacements sur le réseau routier de l'agglomération), et de l'autre côté, face à des textes dans lesquels il s'agit d'argumenter une analyse géopolitique, n'est assurément pas étranger à de telles disparités.

109 Pour une réelle exploitation des titres et, au-delà, de tous les éléments de structuration matérielle des textes (sections, paragraphes, alinéas, etc.), il est nécessaire d'établir une « grammaire » de ces éléments, ou plutôt des grammaires reliées à des genres de texte, grammaires qui seraient des modèles différenciés de structuration des textes.

110 On pourra ainsi tirer parti des indications de mise en forme et de structuration présentes dans certains documents électroniques, à l'instar de la recherche sur le repérage des Entités Nommées que nous avons présentée dans la section 3.

111 Car de plus en plus, les documents numériques comporteront des informations de structuration logique de leur contenu, celles-là mêmes qui peuvent être exploitées dynamiquement pour créer des vues adaptées au lecteur . Cette possibilité de générer, par exemple à travers une feuille de style, une vue particulière, constitue selon nous une métaphore de la fonction de la structuration matérielle des textes : comme le font très justement remarquer , le contenu « n'est appréhendable qu'à travers une mise en forme », il est absolument impossible de le saisir autrement que « sous une forme sémiotique lisible », pour reprendre la formulation des auteurs. En ce sens, tout élément de structuration matérielle contribue à la sémiotisation du contenu et, pour opérer celle-ci, les documents numériques comportent donc tout à la fois ce contenu et un ensemble d'informations qui peuvent être utilisées comme autant **d'instructions** pour la mise en forme de ce contenu. C'est là une richesse considérable pour les futurs systèmes de traitements automatiques, nous revenons dans la conclusion sur ce qu'apporte cette dualité.

6. Conclusion

112 Nous avons avancé un certain nombre d'arguments pour étayer l'idée que la structuration matérielle d'un document peut être mise à contribution pour l'accès à son contenu sémantique. Nous avons commencé par montrer comment l'inscription du texte sur un support matériel peut être porteuse de signification : ce qui est matérialisé par la disposition et les propriétés typographiques peut être considéré comme la contrepartie d'un énoncé discursif dans lequel la signification se construit *via* des formes lexicales et syntaxiques.

113 Ainsi, comme l'indique le Modèle d'Architecture Textuelle, qui fournit un cadre théorique adéquat, les faits de structuration matérielle sont partie prenante de la signification du texte. Il semble donc pertinent d'intégrer la prise en compte de la mise en forme matérielle et des objets textuels qu'elle rend perceptibles à des tâches automatiques visant un accès au contenu des documents, d'autant plus que de nouveaux formats de documents numériques tels xml comportent des informations explicites pour la génération de vues des documents.

114 Cela signifie qu'au lieu de traiter du texte brut, avec souvent des marques de paragraphes ou de fin de ligne pour toute indication de mise en forme, les systèmes de traitements automatiques auront affaire à des données structurées, balisées. Si l'on veut que cette nouvelle richesse soit une réelle plus-value pour l'accès au contenu sémantique des documents, il faut savoir mettre en relation les indications fournies par ce balisage et le contenu sémantique qu'il permet effectivement de produire, autrement dit, il faut savoir l'interpréter, l'inclure dans un calcul du sens pour pouvoir l'exploiter.

115 C'est dans cette perspective que s'inscrit notre analyse des fonctions des titres de section, et que s'ouvre un champ de recherches en linguistique et en traitement automatique des langues.

BIBLIOGRAPHIE

Aït El Mekki T., Nazarenko A. (2002), « Comment aider un auteur à construire l'index d'ouvrage ? L'architecture du système IndDoc », *Colloque International sur la Fouille de Textes (CIFT)*, Tunisie, p. 141-157.

Bachimont B., Crozat S. (2004), « Instrumentation numérique des documents : pour une séparation fonds/forme », *Information-Interaction-Intelligence,* vol. 4(1), p. 95-103.

Bessonnat D. (1988), « Le découpage en paragraphes et ses fonctions », *Pratiques,* vol. 57, p. 81-105.

Biber D. (1988), *Variation Accross Speech and Writing*, Cambridge, Cambridge University Press.

Charlet J., Aït El Mekki T., Bourigault D., Nazarenko A., Teulier R., Toledano B. (2004), « CEDERILIC : constitution d'un livre et d'un index numérique », In P. Enjalbert et M. Gaio (Eds.), *7e Colloque International sur le Document Electronique*, La Rochelle, 22-25 juin, Europia, p. 187-204.

Charolles M. (1997), « L'encadrement du discours : univers, champs, domaines et espaces », *Cahier de Recherche Linguistique,* vol. 6, p. 1-73.

Charolles M. (2002), « Organisation des discours et segmentation des écrits », *Inscription Spatiale du Langage : structures et processus*, Toulouse, 29-30 janvier, Prescot, p. 31-39.

Cornish F. (1998), « Les "chaînes topicales" : leur rôle dans la gestion et la structuration du discours », *Cahiers de Grammaire,* vol. 23, p. 19-40.

Ferret O., Grau B., Minel J.-L., Porhiel S. (2001), « Repérage de structures thématiques dans des textes », *TALN'2001*, Tours, 2-5 juillet, p. 163-172.

Gala Pavia N. (2003), *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires*, Doctorat Nouveau Régime, Université Paris 11.

Goody J. (1979), *La Raison graphique*, Paris, Editions de Minuit.

Ho-Dac M., Jacques M.-P., Rebeyrolle J. (2004), « Sur la fonction discursive des titres », In S. Porhiel et D. Klingler (Eds.), *L'unité texte*, Pleyben, Perspectives, p. 125-152.

Jackiewicz A. (2004), « Les séries linéaires dans le discours : marques, opérations et structures sous-jacentes. » *Journée de l'ATALA : Modéliser et décrire l'organisation discursive à l'heure du document numérique*, La Rochelle, 22-25 juin.

Jackiewicz A., Minel J.-L. (2003), « L'identification des structures discursives engendrées par les cadres organisationnels », *TALN'2003*, vol. 1, Batz-sur-Mer, 11-14 juin, p. 155-164.

Jacquemin C., Bush C. (2000a), "Combining Lexical and Formatting Cues for Named Entity Acquisition from the Web", In H. Schutze (Ed.), *Joint Sigdat Conference On Empirical Methods In Natural Language Processing And Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong.

Jacquemin C., Bush C. (2000b), « Fouille du Web pour la collecte d'Entités Nommées », In E. Wehrli (Ed.), *TALN'2000*, Lausanne, 16-18 Octobre.

Jacques M.-P. (2003), *Approche en discours de la réduction des termes complexes dans les textes spécialisés,* Doctorat Nouveau Régime, Université Toulouse II Le Mirail.

Jacques M.-P., Mojahid M., Sarda L. (2001), « Repérer les structures du texte Eléments pour la construction d'un modèle d'analyse », *Colloque International sur le Document Electronique*, Toulouse, 24-26 Octobre, Europia, p. 99-113.

Laignelet M. (2004). *Les titres et les cadres de discours temporels*. Université Toulouse II le Mirail.

Lambrecht K. (1994), *Information Structure And Sentence Form*, Cambridge, Cambridge University Press.

Luc C. (2000), *Représentation et composition des structures visuelles et rhétoriques du texte. Approche pour la génération de textes formatés,* Doctorat Nouveau Régime, Université Paul Sabatier.

Luc C., Mojahid M., Virbel J. (2001a), « Système notationnel de l'architecture textuelle par image de page », *Colloque International sur le Document Electronique*, Toulouse, 24-26 Octobre, Europia, p. 233-245.

Luc C., Virbel J. (2001b), « Le modèle d'architecture textuelle Fondements et expérimentation", *Verbum,* vol. 23(1), p. 103-123.

Maurel F. (2004), « De l'écrit à l'oral : analyses et générations », *TALN'2004*, vol. 1, Fès, Maroc, 19-22 avril, p. 289-298.

Minel J.-L., Desclés J.-P., Cartier E., Crispino G., Ben Hazez S., Jackiewicz A. (2001), « Résumé automatique par filtrage sémantique d'informations dans des textes », *Technique et Science Informatiques,* vol. 20(3), p. 369-395.

Mitkov R. (1998), "Robust Pronoun Resolution with Limited Knowledge", *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98 / ACL'98)*, Montréal, Canada.

Pascual E. (1991), *Représentation de l'architecture textuelle et génération de texte*, Doctorat Nouveau Régime, Université Paul Sabatier.

Pascual E., Péry-Woodley M.-P. (1997), « Modèles de texte pour la définition », *Actes des 1ères J.S.T. de l'AUPELF-UREF Réseau FRANCIL*, Avignon, p. 137-145.

Péry-Woodley M.-P. (2000), *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*, Toulouse, Equipe de Recherche en Syntaxe et Sémantique.

Rebeyrolle J. (2000), *Forme et fonction de la définition en discours*, Doctorat Nouveau Régime, Université Toulouse II Le Mirail.

Rebeyrolle J., Tanguy L. (2001), « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires", *Cahiers de Grammaire,* vol. 25, p. 153-174.

Virbel J. (1985), « Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle », *Cahiers de grammaire,* vol. 10, p. 5-72.

NOTES

1. La présente recherche s'inscrit dans le cadre du projet Cognitique « *Visualisation dynamique de textes* : extraction sélective, affichage spatial multi-échelle et observation des stratégies de lecture » dont le responsable est C. Jacquemin, voir : pour le texte de présentation du projet.
2. Il semble à peu près assuré que *modèle conceptuel* devrait à l'oral recevoir une intonation spécifique qui signale son statut autonymique : l'expression est employée non pas pour signifier mais pour parler d'elle-même, ce que nous décidons de traduire par la mise en italique.
3. Quelques écrits fondateurs et récapitulatifs de ce modèle sont, dans l'ordre chronologique, . Nous évoquerons dans la prochaine section une exploitation pratique de ce modèle pour le repérage d'éléments spécifiques dans les textes.
4. Une image de page est une représentation des éléments de mise en forme et/ou des informations textuelles pertinents pour l'analyse .
5. Dans , nous avons montré comment s'appuyer sur ce contraste pour désambiguïser plusieurs interprétations possibles de l'architecture textuelle.
6. Notre présentation succincte synthétise les descriptions de , qui ont pour cadre théorique le MAT.
7. Pour l'un des textes des corpus sur lesquels cette recherche se base, pour lequel nous n'avions aucune indication de mise en forme, nous nous sommes appuyées sur la présence d'une numérotation et d'une marque de fin de ligne pour distinguer les titres de section.
8. On en trouvera un exposé plus complet dans .
9. On ne considère pas comme reprise une répétition d'un mot grammatical du titre (préposition, article, ...).
10. La notion de *topic* d'une phrase dénote une relation d'« à-propos »: « The topic of a sentence is the thing which the proposition expressed by the sentence is about. » .
11. Il faut noter que nous employons ici le terme *thématique* dans un sens non technique, pour signifier « l'ensemble des sujets, idées et propositions qui sont développés dans un discours, qui constituent le centre des préoccupations » (à partir de la définition du Petit Robert 96 de *thème* et *thématique*), et non dans le sens de la distinction faite en linguistique entre *thème* et *rhème*.
12. Sur la notion de chaîne topicale, cf.
13. La première partie du titre introduit le « thème » (toujours en un sens non technique, voir note 11) principal de la section : celle-ci concernera le plan appelé Interim Force. La seconde partie introduit un référent – les brigades interarmes intermédiaires – relié à ce thème. Pour mieux comprendre l'extrait, précisons que le texte précédent indique : « le plan "Interim Force" entreprend de mettre sur pied 5, 6 ou même 8 brigades interarmes d'ici à 2007 ».
14. Le découpage en paragraphes a toutefois retenu l'attention de D. Bessonnat .
15. Ce faible effectif interdit l'emploi d'un Chi-deux.

RÉSUMÉS

Nous montrons dans cet article l'articulation de la structure matérielle et du contenu sémantique des documents textuels. Nous défendons l'idée que la caractérisation automatique du contenu textuel bénéficierait d'une meilleure compréhension du rôle de la structure matérielle et qu'il est d'autant plus pertinent de chercher à l'inclure dans des traitements automatiques que cette structure matérielle est explicitée sous les formats qu'adoptent les textes conçus pour des supports numériques (par ex. html ou xml). Nous inscrivons notre recherche dans le cadre du Modèle d'Architecture Textuelle, qui fournit un modèle théorique pour la définition et l'analyse des objets textuels signalés par des propriétés de mise en forme matérielle. Nous nous focalisons plus particulièrement sur l'analyse des fonctions d'un de ces objets textuels, les titres de section, et sur la façon de les exploiter pour un accès automatique au contenu du document.

My aim in this article is to demonstrate that natural language processing, especially automatic analysis of the content of a textual document, may benefit from a greater understanding of how the visual properties of texts intervene in the construction of meaning. It is particularly worth studying the role of layout in meaning insofar as layout features may be explicitly mentioned in xml documents. I adopt the point of view of the model of text architecture, which provides a theoretical framework to analyze layout and visual features. I focus on the analysis of the varied functions of headings in a text and I show how to exploit headings to characterize the content of the section they head.

INDEX

Mots-clés : propriétés matérielles des textes, structuration du discours, Modèle d'Architecture Textuelle, fonctions des titres de section, traitement automatique, navigation intradocumentaire
Keywords : visual properties of texts, discourse structure, model of text architecture, function of headings, natural language processing, intra-document navigation

AUTEUR

MARIE-PAULE JACQUES

ERSS - UMR 5610, Université de Toulouse-Le Mirail