RÉGAL, un système pour la visualisation sélective de documents

Javier Couto, Olivier Ferret, Brigitte Grau, Nicolas Hernandez, Agata Jackiewicz, Jean-Luc Minel, Sylvie Porhiel

▶ **To cite this version:**

Javier Couto, Olivier Ferret, Brigitte Grau, Nicolas Hernandez, Agata Jackiewicz, et al.. RÉGAL, un système pour la visualisation sélective de documents. Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle, 2004, 18 (4), pp.481-514. 10.3166/ria.18.481-514 . hal-00068746

HAL Id: hal-00068746
https://hal.science/hal-00068746

Submitted on 10 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RÉGAL, un système pour la visualisation sélective de documents

Javier Couto[1], Olivier Ferret[2], Brigitte Grau[3], Nicolas Hernandez[3], Agata Jackiewicz[1], Jean-Luc Minel[1], Sylvie Porhiel[1,4]

[1]*Laboratoire LaLICC –Université Paris Sorbonne, CNRS*
96, bd Raspail, 75006 Paris
{prénom.nom}@paris4.sorbonne.fr

[2]*Laboratoire LIC2M - CEA*
18, rue du Panorama,
92265 Fontenay aux Roses Cedex
Olivier.Ferret@cea.fr

[3]*Laboratoire LIMSI - CNRS*
BP133, 91403 Orsay Cedex
{prénom.nom}@limsi.fr

[4] *Université de Chypre*
Chypre
sylvieporhiel@hotmail.com.

RÉSUMÉ. *Les systèmes de recherche d'information renvoient généralement une liste ordonnée de documents, où seuls le titre et parfois un extrait comportant les mots de la requête permettent d'en évaluer la pertinence pour son besoin initial. Ces types de résultat conduisent toujours à devoir consulter de nombreux documents pour réellement trouver des réponses pertinentes. Afin d'éviter cet écueil nous nous sommes intéressés à la visualisation d'un texte : que doit-on montrer et comment ? Dans notre système, RÉGAL (RÉsumé Guidé par les Attentes du Lecteur), les informations nécessaires à la visualisation sont extraites automatiquement des textes par l'application d'une analyse thématique, sans présupposer l'existence d'une structure préalable ou d'un formatage du texte, et en combinant des approches fondées sur la cohésion lexicale et le repérage de marques clés.*

ABSTRACT. *Information retrieval systems generally return a list of ranked documents, such as only the title and possibly a snippet that contains the words of the request allow a user to evaluate the document relevance relative to her initial request. This kind of result leads the user to browse a lot of documents before satisfying her information need. In order to improve information retrieval, we have studied text visualization: which information has to be shown and how? Our system RÉGAL (RÉsumé Guidé par les Attentes du Lecteur), automatically extracts the visualized information from texts by applying a thematic analysis that does not require a pre-existing structuring or a formatting of the texts, and is based on the combination of two criteria: lexical cohesion and cue phrases.*

MOTS-CLÉS : *visualisation de texte, navigation textuelle, résumé dynamique, analyse thématique.*

KEYWORDS: *text visualization, textual navigation, dynamic summarization, topic analysis.*

1. Introduction

Les systèmes de recherche d'information renvoient généralement une liste ordonnée de documents où seuls le titre, et parfois un extrait comportant les mots de la requête, permettent d'en évaluer la pertinence par rapport au besoin initial de l'utilisateur. Ces types de résultats conduisent souvent celui-ci à devoir consulter de nombreux documents pour trouver des réponses véritablement pertinentes. Qui plus est, ces documents ne sont souvent qu'une version électronique d'une présentation papier à laquelle on a éventuellement ajouté quelques liens. De ce fait, un utilisateur n'a pas la possibilité d'évaluer rapidement si le document est pertinent ou non sans le lire dans son intégralité. Cela tient en grande partie au fait que les documents mis à disposition sur un support électronique ne sont pas conçus pour une navigation et une consultation rapide. Ils contiennent au mieux un résumé.

À partir de cette évaluation des systèmes de recherche d'information, nous nous sommes demandés comment rendre rapidement disponible l'information contenue dans un texte, et plus particulièrement l'information recherchée par un utilisateur. Cette perspective nous a conduits à écarter la production d'un résumé statique (Minel 2003 ; Marcu, 1997 et 2000) qui est, par définition, sélectif dans l'information qu'il extrait. Par exemple, un système peut sélectionner des textes à partir d'une requête sur un sujet X bien que le sujet X n'en constitue pas un thème principal ; de ce fait, ce thème risque de ne pas figurer dans le résumé, et l'utilisateur aura malgré tout à lire l'intégralité du texte. Afin d'éviter cet écueil, nous nous sommes intéressés à la visualisation d'un texte : que doit-on montrer et comment ? Cette visualisation doit être à la fois indicative et informative (Saggion *et al.*, 2000), (Kan *et al.,* 2001), (Boguraev *et al.*, 2001). Indicative, pour savoir quels sont les différents thèmes du document ; informative pour donner des renseignements sur les segments du texte qui en parlent : leur taille, leur position dans le texte, à la fois d'un point de vue linéaire mais aussi leur position dans la structure thématique du texte, et enfin donner un accès au contenu des segments eux-mêmes. De plus, en maintenant des liens entre toutes ces entités, on permet de naviguer d'un sujet à l'autre, d'un type d'information à un autre, et on permet ainsi une adaptation dynamique de la visualisation selon les besoins de l'utilisateur.

Dans notre système, RÉGAL (RÉsumé Guidé par les Attentes du Lecteur), les informations nécessaires à la visualisation sont extraites automatiquement des textes, sans présupposer l'existence d'une structure préalable ou d'un formatage du texte. Notre but est de concevoir un modèle générique applicable à différents domaines sans avoir à modéliser de nouvelles connaissances. Toutefois, nous avons limité le genre de document traité. En effet les caractéristiques concernant le vocabulaire et le style peuvent être différentes d'un genre à l'autre (Illouz, 2000) et la méthode de segmentation thématique que nous utilisons, qui est fondée sur la cohésion lexicale, est sensible à ces caractéristiques (Ferret et al., 1998). Nous avons choisi de travailler sur des articles scientifiques ou techniques ainsi que sur certains articles journalistiques, textes dits *expositifs*, par opposition à des textes plus *narratifs*. Le

modèle de structuration de texte que nous avons élaboré répond à deux contraintes : être le plus robuste possible tout en étant le plus précis possible. La robustesse nous a conduit vers une méthode de structuration fondée sur la notion de cohésion lexicale (Halliday *et al.*, 1976), capable de produire des résultats sur des textes relevant de tout domaine. C'est par analyse de la récurrence et de la distribution du vocabulaire que nous segmentons un texte en unités thématiquement homogènes. Afin de rendre cette méthode plus précise, nous avons intégré des contraintes supplémentaires fondées sur l'exploitation de marques linguistiques relatives à l'organisation du discours (Charolles, 1997). Deux familles de marques de nature méta-discursive ont été jugées pertinentes : (i) les marques introductrices de thématiques, (ii) les marques introductrices de séries linéaires. L'identification et l'interprétation de ces marques sont réalisées par la méthode d'exploration contextuelle (Desclés *et al.*, 1997). Cependant, ces marques ne peuvent être les seuls critères de segmentation car d'une part, elles sont assez rares et d'autre part, elles ne permettent pas de délimiter entièrement des segments textuels. Elles marquent souvent des débuts de segment, à l'instar de l'annonceur de thème « En ce qui concerne », alors que la délimitation automatique de leur portée est souvent très difficile, voire impossible. En revanche, elles possèdent un fort degré de fiabilité. De plus, alors que la plupart des systèmes se limitent à une segmentation linéaire du texte (Hearst, 1997 ; Kozima, 1993), notre modèle permet de construire une structure organisant les différents segments repérés. Le critère de regroupement est fondé sur une proximité lexicale.

Les thèmes d'un texte, et plus précisément les thèmes de chaque segment, sont identifiés par des groupes nominaux extraits du texte. Cette extraction repose elle aussi sur des critères de récurrence et de distribution qui permettent de sélectionner les descripteurs les plus caractéristiques d'un segment. La récurrence ne se limite pas dans ce cas à comptabiliser les formes identiques mais intègre également les expressions référant à une même entité afin de tenir compte de la variabilité de la langue.

Après un état de l'art permettant de positionner notre approche, nous donnons une vision globale de RÉGAL en section 3, ce qui permet d'introduire chaque module du système. La section 4 développe les méthodes de segmentation thématique suivant les deux approches choisies et précise comment nous les faisons collaborer, avant de présenter la méthode de structuration. Ensuite, nous présentons l'enrichissement du texte par des annotations sémantiques permettant de caractériser les thèmes et typer certaines phrases (section 5) avant de préciser la mise en œuvre du modèle dans la plate-forme ContextO (section 6). Enfin la section 7 expose la visualisation que nous proposons dans RÉGAL avant de présenter un exemple complet section 8 et de conclure.

2. Systèmes d'accès à l'information textuelle

Comme en témoigne l'intérêt pour les conférences TREC (*Text Retrieval Conference*), l'essentiel des travaux réalisés en recherche d'information textuelle ont été menés dans l'optique de traiter des bases contenant de nombreux documents (Cutting *et al.*, 1992 ; Wise *et al.*, 1995 ; Chen *et al.*, 1998). En revanche, un intérêt relatif a été porté jusqu'à présent sur les techniques d'accès au contenu de documents simples. On retrouve néanmoins les travaux de (Jacquemin *et al.*, 2002) pour la visualisation de larges documents, ainsi que bon nombre de travaux en résumé automatique (Boguraev *et al.*, 1999 ; Saggion *et al.*, 2000 ; Choi, 2000b).

Outre une distinction par la taille et le nombre de documents qu'ils considèrent, ces systèmes sont tous confrontés au même problème majeur : en matière de représentation d'information textuelle il est très difficile de représenter du texte autrement que par du texte. « Le langage est notre principal moyen de communication d'idées abstraites pour lesquelles il n'y a pas de manifestation physique évidente » (Hearst, 1999). La solution d'abstraction la plus communément adoptée, quelles que soient les techniques de visualisation, consiste à organiser et assembler l'information présentée suivant des critères thématiques.

Les systèmes de recherche d'information dans des bases documentaires utilisent des techniques de classification pour regrouper des documents. Leur présentation repose ensuite sur des paradigmes visuels 2D ou 3D (Wise *et al.*, 1995 ; Hearst, 1999). Néanmoins, plus on s'intéresse au traitement de documents individuels et de petites tailles (10 à 20 pages), plus se fait ressentir la difficulté d'abstraire du texte de manière graphique.

La plupart des systèmes de recherche d'information fonctionnant à partir d'une requête, la solution de présentation la plus intuitive du contenu d'un document consiste à présenter le titre du document ou bien à surligner les mots de la requête en contexte (*KWIC*[1]). Mais dans une optique de lecture rapide, les systèmes ne peuvent se limiter à ce genre de technique. Les interfaces de type « TileBars » représentent une avancée vis-à-vis de ces dernières techniques (Hearst, 1996). Une barre graphique est affichée à côté du titre de chaque document retourné. Elle représente les zones du textes qui contiennent tel ou tel terme de la requête. Bien que n'étant pas associée à des moyens de navigation intra-document, cette présentation permet de rapidement cibler les zones pertinentes d'un document pour un thème donné. Il n'est cependant pas toujours évident pour un utilisateur de spécifier à l'avance des patrons de ce qu'il recherche exactement.

Le système SumUM (Saggion *et al.*, 2000) se présente comme un « résumeur » automatique et dynamique de documents scientifiques. Il propose un résumé en deux étapes : d'abord indicatif, en énonçant les thèmes majeurs du document, puis de nature informative en développant un aspect particulier laissé au libre choix du lecteur. La méthode qui le sous-tend repose sur une interprétation des textes par une

[1] *Keyword-in-context.*

analyse sélective de leur contenu à partir de patrons syntaxiques et lexicaux. Un résumé est ensuite engendré en s'appuyant sur un ordonnancement rhétorique préétabli et sur la structure thématique des textes (termes des titres).

Bien que ce système ait obtenu en moyenne les meilleurs résultats lors de la campagne d'évaluation de techniques de résumé automatique de la conférence DUC 2002, il nécessite des connaissances très coûteuses en développement. Dictionnaire conceptuel, identification des types d'information pertinente, définition des patrons de reconnaissance indicatifs et informatifs et du plan de génération ont été ainsi réalisés manuellement par observation comparative de textes et de leur résumé.

(Boguraev *et al.*, 1999) introduisent les notions « d'aperçu local » - segments de textes thématiquement homogènes - et « d'étiquette thématique » - syntagmes les plus pertinents d'un segment (potentiellement enrichis du contexte textuel proche) - afin de proposer une vision facilitant l'accès au contenu d'un texte tout en restant intelligible. La segmentation est obtenue par mesures lexicales tandis que l'identification des termes saillants repose sur des heuristiques linguistiques (prédominance de fonctions grammaticales, de positions privilégiées dans la phrase, récurrence des mots, etc.). L'outil « ViewTool » présenté dans cet article combine des techniques simples de visualisation (entre autres : contexte et focus, aperçu et détails zoomés) sur un écran divisé en 3 parties : la première où le texte est compressé afin d'être entièrement visible, même sous forme schématisée, et fournir ainsi un contexte. La deuxième, réagissant au survol d'une zone du texte de la première, renvoie un aperçu de son contenu à travers les termes saillants. Et la troisième qui occupe plus de la moitié de l'écran et offre le détail de la zone survolée.

Avec une approche similaire, (Choi, 2000b) compare et évalue un certain nombre de techniques de visualisation de textes. Il constate notamment qu'une segmentation thématique linéaire d'un texte fournit une unité plus pertinente que les mots, les phrases ou les paragraphes pour naviguer dans des textes de grandes tailles (livres et magazines). De même, cette segmentation complétée par une compression simplifiée du texte des segments (fournissant un semblant de niveau haut de table des matières) donne de meilleurs résultats qu'une technique de compression télégraphique seule (suppression des mots non informatifs). Dans la continuité de son analyse thématique, il propose la construction d'un arbre thématique comme moyen plus fin pour naviguer au sein d'un document. Ciblant la relation thème/sous-thème, il identifie des relations de dépendance entre les segments à partir d'indices tels que la répétition des termes et la présence de connecteurs.

Notre approche se situe dans la lignée de ces deux derniers travaux tout en s'en démarquant par un certain nombre de points spécifiques que nous présenterons tout au long de l'article. Par ailleurs, chacune des études qui contribuent à l'élaboration du modèle général sera resituée dans son contexte au moment où elle sera présentée.

3. Vision globale de RÉGAL

Le système que nous avons réalisé, RÉGAL (RÉsumé Guidé par les Attentes du Lecteur), suit l'architecture présentée figure 1. Cette architecture fait apparaître les différents modules exposés dans la suite de cet article et permet de les situer les uns par rapport aux autres.

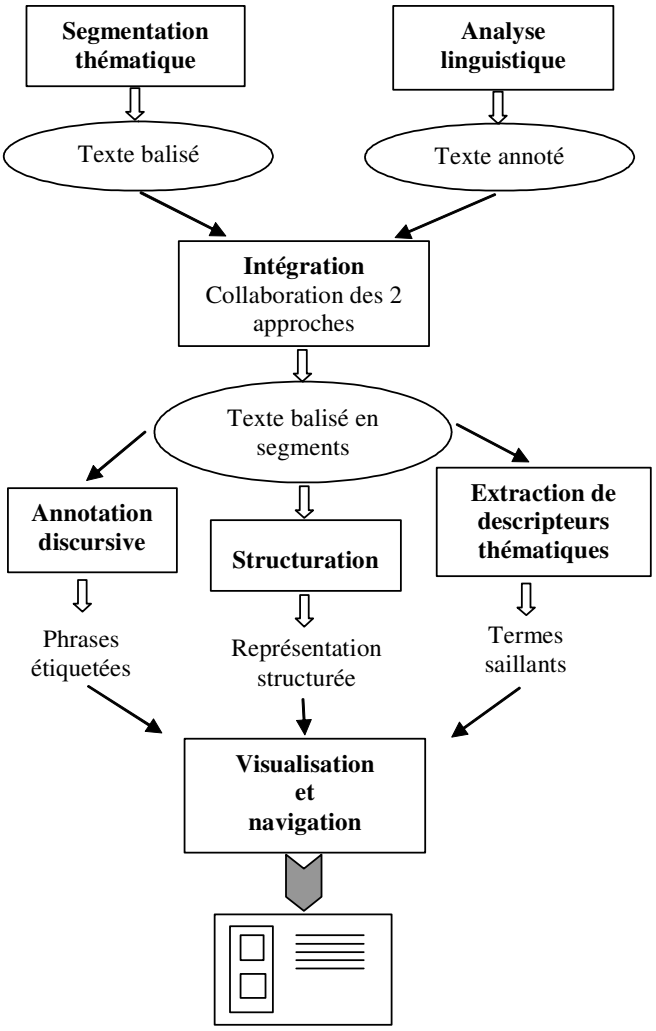

Figure 1. *Le système RÉGAL*

Les deux analyses thématiques du texte ont lieu en parallèle. Le module de segmentation thématique s'appuie sur la cohésion lexicale des textes. Cette segmentation délimite des segments textuels considérés comme fortement cohérents du point de vue du thème qu'ils développent. Les informations de segmentation – début et fin de segment – sont décrites dans un format XML (Hernandez *et al.*, 2002) défini par rapport à une tokenisation de référence du texte considéré. L'analyse linguistique s'appuie pour sa part sur la plate-forme ContextO (Crispino *et al.*, 1999). Elle identifie les débuts de cadre thématique du texte analysé et des informations considérées comme structurantes du point de vue discursif comme les marques d'intégration linéaire. La sortie de ce module est un texte également annoté au format XML. Les annotations sont essentiellement des balises de début de segment textuel.

Le module d'intégration ajuste les segments trouvés en fonction des marques linguistiques présentes. En effet nous considérons ces dernières plus fiables que l'exploitation de la cohésion lexicale. Ces marques signalent en général le début des segments mais plus rarement leur fin. Aussi, en l'absence de marques, nous nous fions aux résultats de la segmentation thématique.

Le module de structuration repose sur le même principe de cohésion lexicale que celui utilisé lors de la segmentation initiale. Chaque segment est décrit par un vecteur de mots et une mesure de similarité permet d'évaluer la proximité entre les segments non contigus. Nous regroupons les segments les plus similaires et considérons que les segments inclus entre ceux-ci sont à un niveau hiérarchique inférieur, ce qui signifie que des thèmes différents ont été introduits lors du développement d'un thème. Cette inclusion peut correspondre à un changement de thème, une digression ou une spécialisation. Notre mode de structuration respecte la linéarité du texte mais seuls les liens entre segments se rapportant à un même sujet sont marqués ; les liens de dépendance entre sujets différents ne sont pas typés.

Afin de caractériser les différents segments de texte et présenter à l'utilisateur des informations significatives, nous rajoutons des annotations sémantiques au texte. Le repérage de certains types d'information tels que les exemples ou les marques de résultats ont déjà fait l'objet de travaux antérieurs (Minel *et al.*, 2001 ; Le Priol, 2000 ; Jackiewicz, 1998) qui sont déjà intégrés dans la plate-forme ContextO. Dans le présent travail, nous avons ajouté des annotations décrivant les thèmes traités dans les segments. Ces descripteurs proviennent d'une extraction automatique de groupes nominaux considérés comme les plus représentatifs des thèmes développés. En procédant ainsi, nous ne produisons aucune généralisation d'un thème si celle-ci n'est pas explicitement mentionnée dans le texte. Ce choix résulte de la volonté de ne dépendre d'aucune source de connaissances extérieure sur les thèmes, une telle source n'existant pas sur une large échelle. Par ailleurs, nous avons préféré laisser à l'utilisateur le soin d'interpréter lui-même les notions développées dans le texte, mais en lui offrant la possibilité de sélectionner facilement les informations nécessaires à cette interprétation.

Enfin, toutes les informations produites par les diverses analyses permettent de proposer une interface de visualisation et de navigation dans le texte. L'utilisateur peut choisir ce qu'il veut voir, se déplacer dans la structure du texte, parcourir les différents types d'annotation ou passer des segments complets à leur description synthétique. Selon ses besoins informatifs, il pourra choisir d'explorer un ou plusieurs aspects du texte.

4. Segmentation et structuration thématique des textes

4.1. Segmentation fondée sur la cohésion lexicale

À l'instar des travaux portant sur l'analyse et la structuration du discours, notre premier objectif est ici de définir des unités textuelles élémentaires caractérisées par une forte cohérence interne. La notion de cohérence est cependant intimement liée à l'interprétation du lecteur et de ce fait, reste encore inaccessible pour l'essentiel aux instruments du traitement automatique des langues. Les travaux de Halliday et Hasan (1976), ainsi que d'autres à leur suite, ont néanmoins montré que la notion de cohésion, plus objectivement caractérisable dans les textes, est un bon indicateur de la cohérence textuelle. La cohésion se manifeste sous plusieurs formes (référence, ellipses, …) mais les travaux relatifs à la segmentation automatique de textes ont particulièrement porté leur attention sur la cohésion lexicale, c'est-à-dire la forme de cohésion résultant de la présence de relations sémantiques entre les mots d'un texte. Plus précisément, ces travaux font l'hypothèse que les ruptures observées dans la cohésion lexicale sont représentatives des variations de cohérence résultant du passage d'une unité textuelle à une autre.

Halliday et Hasan (1976) ont montré que la cohésion lexicale repose sur deux types de relations : les relations de réitération et les relations de collocation. Les relations de réitération correspondent à la reprise d'une entité à l'identique (par le même mot) ou sous une forme plus générale (hyperonyme) tandis que les relations de collocation caractérisent une reprise par le biais d'un concept lié (méronyme, antonyme ou concept appartenant au même thème). La distinction opérée entre ces deux types de relations s'est traduite au niveau des systèmes de segmentation de textes par la distinction de deux types de systèmes. Les plus répandus (Hearst, 1997 ; Choi, 2000a) s'appuient sur la forme la plus minimaliste des relations de réitération : la stricte récurrence lexicale. Ceux exploitant les relations de collocation (Kozima, 1993 ; Ferret, 2002) ont besoin, quant à eux, d'une source de connaissances externe aux textes dans laquelle figurent les relations de collocation nécessaires à l'identification de la cohésion. Comme Ferret *et al.* (1998) l'ont mis en évidence, ces deux types de systèmes sont complémentaires : la seule récurrence lexicale est un indicateur peu exigeant en termes de ressources à mettre en œuvre mais il n'est fiable que si la façon dont un concept est exprimé ne varie pas trop au sein du même texte, comme dans les textes techniques par exemple. Dans le cas contraire, seules les relations présentes dans une source de connaissances permettent d'identifier les liens

de cohésion. L'utilisation de ces relations doit néanmoins rester limitée aux types de textes pour lesquels la récurrence lexicale est peu efficace car une mauvaise représentation du vocabulaire d'un texte dans la source de connaissances peut conduire à privilégier des liens entre des mots qui ne sont pas représentatifs de la thématique de ce texte.

Afin d'assurer au module de segmentation utilisé le champ d'application le plus large possible, nous avons opté pour une méthode s'appuyant sur la récurrence lexicale mais pouvant intégrer dans le même cadre la prise en compte de relations de collocation issues d'une source de connaissances. Nous n'en donnerons ici que les caractéristiques principales, les détails pouvant être trouvés dans (Ferret *et al.*, 1998). La segmentation réalisée par cette méthode prend comme point de départ la structuration des textes en paragraphes. Même si sa valeur en tant qu'unité textuelle peut être ambiguë, le paragraphe est en effet une unité présentant davantage de fondement que des fenêtres de taille arbitraire (cf. (Hearst, 1997) par exemple). L'objectif du module de segmentation est donc de regrouper les paragraphes contigus susceptibles de former une unité homogène du point de vue de leur contenu.

À l'instar du modèle Vector Space utilisé en recherche d'information, chaque paragraphe est transformé en un vecteur de descripteurs. Les descripteurs sont en l'occurrence les formes normalisées des mots pleins des textes (c'est-à-dire les noms, les verbes et les adjectifs.), obtenues à la suite d'un prétraitement morpho-syntaxique. Le poids de chacun des descripteurs au sein du vecteur représentant un paragraphe est nul si le descripteur n'est pas présent dans le paragraphe et inversement proportionnel à la répartition du descripteur parmi les paragraphes du texte sinon. On considère en effet qu'un descripteur est d'autant moins discriminant pour juger de la proximité de paragraphes qu'il est assez uniformément présent dans le texte. Chaque paragraphe ayant été transformé en un vecteur, la similarité entre paragraphes adjacents est évaluée pour chaque couple possible grâce à une mesure vectorielle, en l'occurrence le coefficient de Dice.

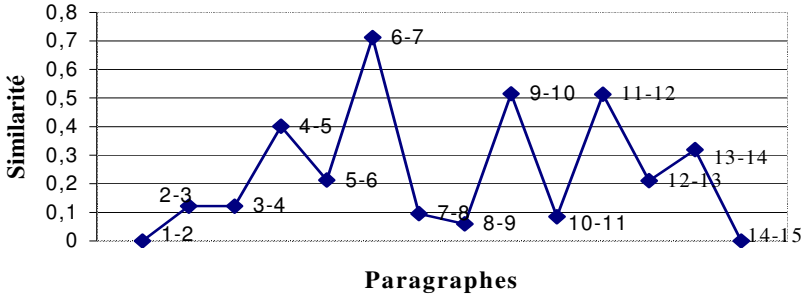

Figure 2. *Courbe de similarité inter-paragraphe pour l'article du Monde de la figure 6*

À l'issue de cette évaluation, on obtient une courbe telle que celle de la figure 2. Les fortes valeurs de similarité, par exemple entre les paragraphes 6 et 7, indiquent des paragraphes proches quant à leur contenu et supposés appartenir à la même unité textuelle. En revanche, les zones de faible similarité, comme entre les paragraphes 7 et 8 ou entre les paragraphes 8 et 9, sont indicatrices d'un changement de contenu et sont considérées comme marquant une frontière d'unité textuelle. En pratique, la détection d'un changement d'unité textuelle est réalisée en comparant les valeurs de similarité obtenues à un seuil dynamique[2] : une frontière est fixée à chaque frontière inter-paragraphe dont la valeur de similarité est inférieure à ce seuil.

La méthode décrite ci-dessus a été adaptée afin de pouvoir y intégrer une source de connaissances externe. Les relations inter-lexicales issues de cette source permettent alors de détecter des similarités entre paragraphes faisant référence au même contenu alors qu'ils n'ont que peu de descripteurs en commun. Dans le cas présent, la source de connaissances utilisée est un réseau de cooccurrences lexicales, c'est-à-dire un ensemble de cooccurrences lexicales recueillies à partir d'un vaste corpus. Les modalités de recueil de ces cooccurrences (Ferret, 2002) ont été fixées afin de favoriser les cooccurrences sous-tendues par des relations sémantiques. La prise en compte de ces cooccurrences est faite de la façon suivante : soient deux paragraphes adjacents ; lorsque deux descripteurs sont fortement liés dans le réseau de cooccurrences et que l'un au moins des deux est présent dans l'un des deux paragraphes, leur poids est renforcé dans le vecteur représentant chacun des deux paragraphes. Dans le cas où le poids du descripteur est initialement nul, cela se traduit donc par un ajout du descripteur dans le paragraphe. Après l'application de cette procédure, la similarité entre paragraphes est évaluée comme précédemment.

Plus globalement, le module de segmentation fondée sur la cohésion lexicale fonctionne en appliquant la première ou la seconde méthode en fonction du type de texte rencontré : si le vocabulaire du texte est peu représenté dans le réseau de cooccurrences lexicales, le texte est présumé de nature technique[3] et seule la récurrence lexicale est utilisée ; dans le cas contraire, les relations du réseau de cooccurrences lexicales sont exploitées.

4.2 Approche linguistique

La précédente approche se fonde sur les relations lexicales que les mots entretiennent pour établir ou non des ruptures thématiques dans les textes. Ces ruptures lexicales se trouvent le plus souvent employées en parallèle avec d'autres procédés cohésifs, qui les renforcent, tels les cadres de discours. Ceux-ci contribuent à partitionner l'information dans des blocs sémantiquement homogènes, en désignant

[2] Ce seuil est défini par : moyenne des valeurs de similarité – α ● écart-type des valeurs de similarité.

[3] Notre réseau de cooccurrences lexicales a en effet été construit à partir du journal *Le Monde* et peut être considéré comme représentatif du vocabulaire « général ».

les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements. M. Charolles (1997) distingue quatre familles de cadres : (i) les cadres vérifonctionnels, dont font partie les cadres spatiaux (*À Chypre*), les cadres temporels *(En 2003)*, les cadres médiatifs (*Selon Piaget*), les cadres représentatifs (*Dans le film de Godard*), les cadres praxéologiques (*En chimie*) ; (ii) les cadres thématiques (*En ce qui concerne la partition de Chypre*) ; (iii) les cadres qualitatifs (*Par chance*) ; (iv) les cadres organisationnels (*En premier lieu*).

Les cadres de discours instaurent un lien de cohésion textuelle que le lecteur reconstruit à partir de nombreux indices linguistiques et en particulier en s'appuyant sur les introducteurs de cadres. La portée de ces marques, syntaxiquement non intégrées à l'énoncé où elles figurent matériellement, généralement en position initiale, peut s'étendre sur plusieurs phrases (voire paragraphes, pour certaines d'entre elles) créant ainsi une véritable unité textuelle, homogène sémantiquement et relativement autonome par rapport au contexte.

Les cadres thématiques et les cadres organisationnels, que nous avons plus particulièrement étudiés dans ce projet, ont en commun deux propriétés essentielles : (i) leurs introducteurs explicitent l'organisation du contenu textuel et sont donc de nature méta-discursive, (ii) leurs introducteurs peuvent fonctionner de concert pour baliser une série de segments discursifs[4].

4.2.1. *Les cadres thématiques*

Les introducteurs thématiques constituent une classe cohésive dont les éléments, de nature abstraite, sont morphologiquement des prépositions (*en ce qui concerne, pour ce qui est de, à propos de, sur*, etc.) et des anaphores résomptives (*à ce sujet, à ce propos*) (Porhiel, 2003). Ces unités lexicales assurent une cohérence thématique, le plus souvent au niveau local dans un texte. Nous entendons par là qu'un texte est rarement entièrement structuré par des introducteurs thématiques ce qui, en revanche, peut être le cas avec des introducteurs de cadres spatiaux ou des introducteurs de cadres temporels.

Au niveau textuel, les introducteurs thématiques servent à introduire un élément dont il va être question dans la proposition et les phrases subséquentes : ils en précisent le thème. Ce sont des balises linguistiques identifiables dans tout type de texte : elles pointent sur une thématique et renforcent une segmentation thématique déjà opérée au niveau lexical (pronoms, répétitions, synonymie) et/ou orthographique (marques de paragraphes). Les introducteurs ont aussi pour fonction de séquencer explicitement des parties d'un texte : ils attirent l'attention sur un référent et le rendent saillant par rapport à d'autres choix possibles ; ils organisent l'information dans un texte et la répartissent dans des blocs sémantiquement homogènes et ont, dans ce sens, une « portée étendue ».

[4] Il s'agit d'une propriété inhérente aux cadres organisationnels, ce qui n'est pas le cas pour les cadres thématiques.

La répartition de l'information se fait de deux façons : soit il n'y a qu'une seule occurrence d'introducteur thématique, auquel cas, ce dernier fournit le cadre de ce qui va être traité ; soit il apparaît dans une liste (une série) dont le fonctionnement se rapproche de celui des marqueurs d'intégration linéaire. On remarque alors la présence d'une amorce (*le programme industriel…*) puis d'une série d'items dont l'ordre peut être interverti, sauf dans le cas du marqueur *quant à* dont le rôle est de clore ou de servir de relais dans une énumération (Fløttum 1999) :

1. *[§] Le programme industriel doit comporter des définitions précises portant sur les biens prioritaires et sur le comportement des agents.* ***En ce qui concerne*** *les biens, les priorités devraient être établies compte tenu du niveau moyen de revenu et du capital par tête*

En ce qui concerne *les agents, il faut tenir compte de l'élargissement considérable des fonctions de l'État au cours des quinze dernières années* (Le Monde Diplomatique)

Les unités lexicales qui introduisent potentiellement des cadres thématiques sont pour la plupart des prépositions : elles sont donc intrinsèquement polycatégorielles. Pour pouvoir être utilisées dans la plateforme ContextO, elles ont fait l'objet d'une analyse linguistique approfondie. Les informations linguistiques ci-dessous ont été capitalisées en vue du repérage des introducteurs dans des textes (Porhiel, 2003) :

- dans la base de connaissances, la déclaration des formes des marqueurs doit tenir compte de la réalité discursive et linguistique. Ainsi, certains marqueurs se mettent au pluriel (*au chapitre de, aux chapitres de* ; *au sujet de, *aux sujets de*), d'autres varient en temps (*en ce qui concerne, en ce qui concernait*), d'autres encore ont une forme résomptive (*au sujet de, à ce sujet ; aux niveau de, *à ce niveau*). À ce propos, il faut distinguer entre les possibilités linguistiques et les possibilités discursives : quelles que soient les variations que les marqueurs subissent, elles ne doivent pas modifier le fait que ce sont des introducteurs thématiques.

- les introducteurs thématiques sont de nature abstraite et acceptent, en termes de compatibilités lexicale et sémantique, tout type de complément. Toutefois, comme les prépositions n'instaurent pas de relation unilatérale, une même préposition peut potentiellement être un introducteur de cadre (2) ou un complément circonstanciel (3) :

2. Au niveau des *individus, citoyens et consommateurs, l'éducation, l'information, la prise de conscience, l'affirmation de la dimension éthique doivent contribuer à faire évoluer les systèmes de valeurs et les comportements, (…) (*Le Monde Diplomatique*)*

3. Au niveau de *la rétine, la densité des cellules à bâtonnet - qui permettent la vision en noir et blanc avec des éclairages faibles - est beaucoup plus réduite que chez l'adulte. (*Le Point*,* 18 janvier 1997*)*

Il a donc fallu inhiber les possibilités relationnelles non abstraites des prépositions, telles les relations spatiales ou avec une partie du corps avec *au niveau de*.

- en outre, une même unité lexicale peut potentiellement être un introducteur (*Concernant le résultat…*) ou une conjonction limitant sémantiquement un autre constituant morphosyntaxique (*concernant les textes de loi…*). Comme un introducteur thématique préfixe, prototypiquement, au moins une proposition et au plus un paragraphe, la différence entre une expression introductrice de cadre et une conjonction se fait en termes de dépendance, ce qui se traduit ici en termes positionnel : l'introducteur thématique d'une phrase se trouve en position initiale.

> ***4. Concernant*** *le résultat proprement dit, je crois qu'il conforte ma position. J'ai toujours dit que la question de la réduction du mandat de sénateur n'était pas taboue. Mais, pour ma part, je souhaite qu'une telle réduction, si elle était décidée, s'accompagne d'un renforcement des pouvoirs du Sénat* ***concernant*** *les textes de loi relatifs aux collectivités territoriales. D'ailleurs, j'ai déposé une proposition de loi pour réviser la Constitution et allant dans ce sens. (*Le Point*, 6 octobre 2000)*

- les introducteurs composés acceptent des insertions adverbiales et adjectivales (elles sont limitées à 1, 2 ou 3 mots). Il est important de les prendre en compte, du fait de l'emploi combiné des introducteurs thématiques et des marqueurs d'intégration linéaire dans des séries.

- enfin, certains groupes de mots peuvent se placer avant les introducteurs thématiques. Tout comme ces derniers, ils sont en position détachée. De longueur variable, ils sont ou non séparés par une virgule.

> ***5.*** *Toujours dans le numéro de mars, mais* ***à propos de*** *l'article de Christian de Brie, « Voyage au coeur des laboratoires du Front national », un lecteur de Marseille, M. Serge Bonnefoi, précise que (…) (*Le Monde Diplomatique*)*

La capitalisation de ces informations linguistiques a permis de déclarer 122 unités linguistiques, réparties en 21 classes en fonction de leurs caractéristiques morpho-syntaxique et lexico-sémantique. Dans la plateforme ContextO ils constituent des indicateurs qui, combinés à des indices secondaires lexico-sémantiques, positionnels et syntaxiques déclenchent quatre types de règles : l'introducteur est en position initiale, l'introducteur se trouve après a) un groupe, b) deux groupes ou c) trois groupes de mots. Ces règles ont pour fonction d'attribuer à la phrase qui ouvre le cadre thématique une étiquette qui permet de typer cette phrase (cf. 4.2.). Il sera prochainement possible de relier les différents items d'une série ainsi que de prendre en compte le fait que les cadres thématiques et les cadres organisationnels peuvent être combinés dans une même série.

4.2.2. *Les marqueurs d'intégration linéaire comme introducteurs de séries dans le discours*

Les marqueurs d'intégration linéaire (MIL) ont la propriété caractéristique d'être indépendants des contenus sémantiques des segments textuels qu'ils introduisent et relient entre eux sur le mode d'une série linéaire (*d'une part… de l'autre…* ; *en premier lieu … en second lieu…* ; *premièrement… deuxièmement… troisièmement…*). Chaque marqueur d'intégration linéaire constitue à lui tout seul un introducteur de cadre à même d'indexer plusieurs propositions, voire paragraphes. Chaque MIL introduit un cadre organisationnel ; l'ensemble des MIL qui fonctionnent de concert introduisent une structure organisationnelle composée d'une série de segments textuels. Cette structure en série est généralement introduite par une amorce, c'est-à-dire une annonce qui explicite le principe fédérateur des items de la série et en précise la longueur (Jackiewicz, 2002).

> *Quatre-vingts ans après, l'Union soviétique a fait naufrage, et le monde connaît une nouvelle grande mutation, que nous pourrions appeler la seconde révolution capitaliste. Elle résulte, comme la première, de la convergence d'un faisceau de transformations survenues dans trois champs.*
>
> ***En premier lieu**, dans le domaine technologique. L'informatisation de tous les secteurs d'activités ainsi que le passage au numérique (...) bouleversent le travail, l'éducation, les loisirs, etc.*
>
> ***En deuxième lieu**, dans le domaine économique. Les nouvelles technologies favorisent l'expansion de la sphère financière. Elles stimulent les activités possédant quatre qualités: planétaire, permanente, immédiate et immatérielle. (…)*
>
> ***En troisième lieu**, dans le domaine sociologique. Les deux bouleversements précédents mettent à mal les prérogatives traditionnelles de l'État-nation et ruinent une certaine conception de la représentation politique et du pouvoir*. (…) (Le Monde Diplomatique)

La longueur des séries balisées par les marqueurs d'intégration linéaire varie typiquement entre deux et dix éléments. 75% de ces séries sont composées de deux ou de trois segments, ce qui nous renseigne sur la longueur typique de ces séries. Ce résultat dévoile à son tour une propriété saillante des séries courtes, à savoir l'hétérogénéité de leurs introducteurs. Cette caractéristique tient au fait que les introducteurs particuliers issus des séries d'origine temporelle, spatiale ou « numérique » peuvent se combiner entre eux (exemples : *premièrement / en deuxième lieu / enfin* ou *tout d'abord / deuxième facteur / en troisième lieu / enfin, dernier élément)*. Notons au passage que le balisage des séries peut également être incomplet, avec un ou deux introducteurs manquants. Inversement, plus une série est longue, plus ses introducteurs tendent à être homogènes (appartenir à la même série originelle), ce qui explique le très faible nombre de possibilités effectivement exploitées pour introduire une série longue.

La mise en texte d'une série dans le discours n'appelle pas de disposition visuelle spécifique. Plusieurs cas de figure sont possibles, allant du paragraphe compact qui contient à la fois le segment amorce et l'ensemble des éléments de la série s'enchaînant linéairement, à une suite parallèle de paragraphes nettement distincts visuellement, introduits chacun par un MIL en position initiale et par une marque graphique (tiret, puce, numéro). Le plus souvent, les items de la série appartiennent à un seul niveau, contrairement aux énumérations structurées par des marques typographiques et dispositionnelles, pouvant présenter plusieurs niveaux d'enchâssement.

L'idée fédératrice qui relie entre eux les segments d'une série est généralement explicitée dans un segment textuel (ou amorce) précédant la série. Parmi les marques qui renvoient au principe fédérateur de la série, il y a des classifieurs *(éléments, étapes, causes…).* Ces éléments, qui ne peuvent être définis qu'en extension, sont souvent repris au niveau de l'expression cadrative *(elle pourrait être précipitée par trois éléments : premièrement… deuxième élément… troisième élément…).* Le segment amorce donne également une indication sur le nombre d'items de la série. La longueur peut être indiquée (i) précisément (*deux, double…*) ou d'une manière approximative (*plus d'un*, *nombreux…*), ces deux types de marques se combinant avec le classifieur, (ii) ou plus implicitement, par des expressions telles que *un paradoxe, une dichotomie…* Il n'existe pas toujours d'annonce explicite, mais la séquence textuelle qui précède la série est réellement indispensable à son interprétation (par exemple, quand elle exprime une thèse que les différents items de la série ont pour charge d'étayer).

La fermeture du dernier cadre organisationnel de la série, comme c'est le cas de tous les cadres de discours, n'est pas marquée explicitement. Des complexes de plusieurs indices (de nature thématique, graphique...) en général liés à l'ouverture d'une nouvelle structure, amènent le lecteur à fermer ce dernier segment et clore ainsi toute la série. Nous avons constaté toutefois que les séries discursives étaient fréquemment suivies par une évaluation rétrospective portant sur l'ensemble de leurs items ; cette évaluation est introduite par des marques comme (*l'ensemble, les deux, tous…*).

Les connaissances relatives aux séries de cadres organisationnels capitalisées dans la base ContextO se répartissent dans deux grands ensembles. Le premier est représenté par les introducteurs des items formant à l'origine 182 séries différentes. Dans chacune de ces séries, le MIL d'ouverture (*d'abord, premièrement…*) est considéré comme un indicateur, les suivants ont le statut d'indice complémentaire. L'organisation informatique des MIL de rang supérieur ou égal à deux (MIL_2, MIL_3…) est fondée sur la constitution de paradigmes (classes) dont les éléments sont tous les MIL pouvant occuper respectivement le rang 2, le rang 3… dans la série ouverte par le MIL de rang 1 (indicateur). Cette organisation, illustrée dans le tableau 1 pour le marqueur d'abord, augmente considérablement la combinatoire des MIL pouvant s'enchaîner dans une série, sans pour autant s'autoriser à combiner

librement tous les MIL_1 avec tous les MIL_2… jusqu'au MIL_n (avec n=10). L'union de l'indicateur et des indices constitue la signature de la série.

MIL1 indicateur	MIL2	MIL3	MIL4	MIL5
&d_abord	&d_abord2	&d_abord3	&d_abord4	&d_abord5
d'abord	*d'un autre côté, dans un second temps, …*	*dans un troisième temps, le troisième, …*	*quatrièmement, quatrième, …*	*cinquième*

Tableau 1. *Extrait de la table des MIL pour les séries ouvertes par d'abord*

Le deuxième ensemble réunit les données impliquées dans la désambiguïsation des formes, ainsi que celles qui participent à l'identification de l'énoncé introducteur. Tous ces marqueurs font partie des indices complémentaires.

Les indices qui interviennent dans l'interprétation d'une occurrence de forme en tant que marqueur d'intégration linéaire ont trait à la position et aux éléments pouvant figurer dans le co-texte immédiat d'un MIL. Le critère de position spécifie qu'un MIL peut figurer soit à l'initiale de la phrase, soit dans une incise délimitée par deux virgules. Si ce critère n'est pas vérifié, on considère la nature de l'élément qui précède immédiatement le MIL. Parmi les indices autorisés à figurer dans le contexte gauche d'un MIL citons (i) les signes de ponctuation tels que deux-points ou point-virgule, (ii) certains connecteurs (*mais, et, surtout…*) ou marqueurs de clôture (*enfin, en dernier lieu*), (iii) divers marqueurs appartenant à la liste {*si, c'est…*}, (iv) les marques graphiques d'énumération telles que tirets, puces, numéros.

Le traitement de l'énoncé introducteur dans sa forme explicite s'appuie sur l'identification de ses éléments saillants : le classifieur et le cardinal (indiquant la longueur de la liste). Notre liste des classifieurs comprend 172 formes. Les indicateurs de longueur sont au nombre de 50; ils sont classés en deux sous-ensembles selon leur aptitude à se combiner (*deux* éléments) ou non (*une dichotomie*) avec le classifieur. En l'absence de ces marques (le cas d'une amorce implicite), le système sélectionne la phrase qui précède la phrase d'accueil du MIL d'ouverture.

Le processus d'identification de l'ensemble de la série des cadres organisationnels fait appel à plusieurs règles heuristiques qui doivent respecter les contraintes imposées par le modèle sous-jacent à cette structure. Ainsi, le repérage de l'indicateur d'une série déclenche le processus de recherche des indices situés nécessairement en aval. L'étendue de cette recherche est bornée par l'observation empirique que la portée d'un MIL n'excède généralement pas deux paragraphes.

Comme dans le cas des cadres thématiques décrits précédemment, l'application des règles a pour premier résultat l'attribution d'une étiquette aux phrases qui contiennent un introducteur de cadre organisationnel. Dans un deuxième temps, du fait de l'organisation en série des cadres organisationnels, une structure de données spécifique (Jackiewicz *et al.*, 2003) décrivant tous les éléments d'une série est créée et vient enrichir le modèle du texte (Crispino *et al.,* 1999).

4.3. Intégration

La plupart des travaux existants en segmentation considèrent isolément les différents types d'indicateurs de segmentation. On trouve néanmoins quelques usages de combinaisons de marques. (Passonneau *et al.*, 1997), ayant observé la combinaison de pauses, de mots-clefs et de chaînes lexicales co-référentielles pour la segmentation de textes, constatent que le recoupement de marques donne une meilleure précision que les marques prises isolément. Toujours dans un objectif de structuration de textes, (Marcu, 1997) s'intéresse aussi à la structuration par connecteurs. Il signale néanmoins qu'en leur absence, la cohésion lexicale lui permet de lier ou non des segments contigus.

Segmentation par cohésion lexicale	**Analyse linguistique**	**Intégration**
<text> <segCoLex 1> 1 _____ 2 _____ 3 _____ </segCoLex 1> <segCoLex 2> 4 _____ 5 _____ </segCoLex 2> <segCoLex 3> 6 _____ 7 _____ 8 _____ 9 _____ </segCoLex 3> </text> NB : les chiffres sont les n° de ligne	<text> 1 <intrThem 1>_____ 2 _____ 3 <MIL 1/> _____ 4 <MIL 2/> _____ 5 <MIL 3/> _____ 6 _____ 7 _____ 8 <intrThem 2> _____ 9 _____ </text>	<text> <segCoLex 1> 1 _____ 2 _____ </segCoLex 1> <segCoLex 2> 3 <MIL 1/>_____ 4 <MIL 2/>_____ 5 <MIL 3/>_____ </segCoLex 2> <segCoLex 3> 6 _____ 7 _____ </segCoLex 3> <segCoLex 4> 8 _____ 9 _____ </segCoLex 4> </text>

Figure 3. *Exemple d'intégration des résultats des deux analyses thématiques*

Notre approche s'inspire de ces travaux. La segmentation par cohésion lexicale ainsi que le repérage des cadres thématiques et des marqueurs d'intégration linéaire constituent des analyses visant à repérer des marques de structuration du discours. En effet, les segments obtenus par cohésion lexicale fournissent successivement une marque d'ouverture et de fermeture de segments (<segCoLex i>, </segCoLex i> dans la figure 3), les introducteurs thématiques se présentent comme des marques d'ouverture (<intrThem i/>) et les séries de MIL (<MIL i/>) comme une suite de marques d'ouverture et de fermeture, exceptée pour la dernière.

L'étude de différents textes nous a conduits à faire l'hypothèse que ces marques participent toutes à la détermination de la même structure de texte (Ferret, *et al.,* 2001). En considérant les informations provenant des marques linguistiques comme prépondérantes, nous avons retenu le principe d'intégration illustré figure 3. Les première et deuxième colonnes donnent des exemples de résultats obtenus par les différentes analyses. La dernière représente le résultat après intégration des segmentations obtenues. Les numéros représentent les phrases du texte.

Soit l'algorithme d'intégration suivant :

1. Prétraitement : ajustement des marques obtenues lors de la segmentation par cohésion lexicale sur les marques linguistiques voisines à une phrase près. Ce réalignement se justifie par le fait que l'analyse linguistique est considérée comme portant un regard plus fin sur le texte que l'analyse statistique lexicale.
2. Par défaut, toute marque détectée entraîne une segmentation.

   Pour chaque phrase du texte, il y a toujours un segment obtenu par cohésion lexicale en cours, d'où, selon la marque linguistique que l'on détecte :

   a. s'il s'agit d'une MIL, la première d'une série entraîne une estimation de la taille des segments engendrés si l'on segmente à chacun de ses items.

      i. dans le cas où la taille des segments est trop petite, il s'agit d'une énumération locale. On considère alors que tout autre marque ne sera pas prise en compte sur l'étendue de la série, ce qui revient à intégrer les MIL dans un même segment. Ainsi, la taille d'un segment est forcée à au moins deux phrases. Ce paramètre est à rapprocher de la taille de bloc minimal posée par Hearst (1997).

   b. s'il n'y a pas de MIL en cours

      i. soit l'on constate un renforcement par le fait que des marques coïncident.

ii. soit, comme annoncé par défaut, on considère que chaque marque détectée est un indice de rupture thématique suffisant et on segmente.

Lorsque l'on applique l'algorithme aux analyses décrites à la figure 3, on observe les cas suivants :

1) les marques coïncident (e.g. segCoLex 1 et intrThem 1) : il y a renforcement du début d'un segment ;

2) une marque de début (e.g. intrThem 2) est interne à un segment (e.g. segment segCoLex 3 comportant les phrases 6 à 9) : on divise en 2 segments ;

3) il existe des marques d'intégration linéaire à cheval sur 2 segments (e.g. MIL 1 à 3 réparties sur les segments segCoLex 1 et 2) : on réaligne par intégration de toutes les marques dans le même segment (celui qui contient la plus grande partie de l'énumération, segment segCoLex 2 dans l'exemple), ou on segmente à chaque MIL. Le choix dépend de la taille des segments obtenus après segmentation de façon à ce qu'ils ne soient pas trop petits en nombre de phrases.

La différenciation entre le cas 2 et le cas 1 s'effectue à l'issue de l'étape de prétraitement, c'est-à-dire en tenant compte du réalignement des segments fondés sur la cohésion lexicale avec les phrases.

4.4. Structuration

Peu de travaux portent sur l'explicitation automatique de la structure thématique d'un texte sans se fonder sur sa structure logique, c'est-à-dire son découpage préalable en sections ou paragraphes. Certains requièrent une analyse conceptuelle complète des phrases du texte, comme par exemple (Hobbs *et al.*, 1993) ou (Asher *et al.*, 1994) qui explicitent des relations causales et temporelles entre phrases, ce qui est irréalisable pour traiter des textes dans des domaines non restreints. D'autres ne proposent pas une théorie suffisamment bien spécifiée pour envisager sa mise en œuvre ; il en est ainsi des modèles de (Grosz *et al.*, 1986) quant à l'explicitation des intentions dans le discours et (Mann *et al.*, 1988) avec la Rhetorical Structure Theory (RST). (Marcu, 1997) propose la construction automatique de la structure d'un texte, fondée en partie sur la RST. Les relations sont explicitées par l'utilisation d'indices de surface, aussi bien de nature linguistique que fondés sur la cohésion lexicale. Cependant, quelles que soient les marques utilisées, Marcu met en relation des éléments consécutifs et construit une structure arborescente. Il en est de même dans (Yaari, 1997), qui rapproche aussi des segments consécutifs sur un critère de similarité lexicale. Ces approches répondent à une vision « bottom-up » de la structuration en regroupant des unités proches et en les agglomérant. (Salton *et al.*, 1996) proposent, quant à eux, une mesure de similarité lexicale pour construire un graphe des relations entre segments afin de mettre en évidence les thèmes d'un texte. Ceux-ci sont définis comme certains chemins caractéristiques dans le graphe pouvant relier des segments discontinus. Il y

a alors omission complète de la structure linéaire du texte. Seuls les thèmes semblables sont liés, sans qu'ils soient positionnés les uns par rapport aux autres.

Notre objectif n'est pas seulement d'indiquer à l'utilisateur les thèmes traités. Nous voulons aussi présenter la partie du texte qui les développe, en positionnant celle-ci par rapport aux autres parties et en conservant l'organisation spatiale d'origine. À cette fin, nous posons l'hypothèse que l'auteur d'un texte a organisé son propos en suivant une logique d'imbrication et non d'entrelacement des thèmes, cas le plus fréquent pour des articles scientifiques. Nous supposons ainsi que si l'auteur revient à un sujet préalablement traité dans le texte, il clôt la digression ou le changement de sujet en cours et qu'il n'y reviendra pas par la suite. C'est ce principe qui conduit à emboîter des thèmes : nous considérons que les thèmes emboîtés sont secondaires par rapport au thème englobant.

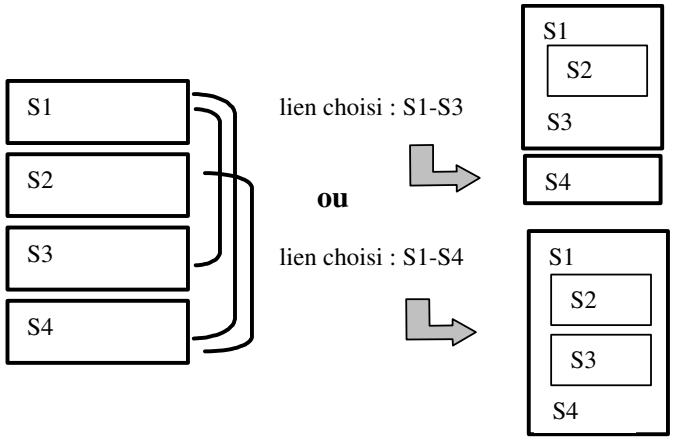

Figure 4. *Exemples du principe d'emboîtement*

Afin de construire la structure emboîtée d'un texte, nous cherchons à mettre en évidence d'abord le niveau le plus englobant en recherchant les deux segments non consécutifs les plus liés. On délimite ainsi la portée du thème développé dans ces passages. Nous ré-appliquons récursivement le même principe aux segments inclus dans la structure de plus haut niveau ainsi qu'à ceux qui sont restés au même niveau. La figure 4 illustre ce principe. Si les segments S1 et S3 sont les plus proches, alors le texte sera structuré en 2 thèmes de même niveau, le premier conduisant à un thème englobé. Si le lien S1-S4 est le plus fort, cela conduit à un seul thème principal, menant à la description de deux aspects différents.

La mesure de liaison de deux segments est la même que celle que nous appliquons pour segmenter le texte et qui est appliquée sur les unités de base. Un segment étant constitué du regroupement de plusieurs unités consécutives, ses descripteurs sont l'union des descripteurs de ces unités (i.e. l'ensemble de tous les

mots retenus), leur poids étant la moyenne des poids de chaque unité de base. Nous avons choisi la moyenne des valeurs afin de conserver des éléments comparables et ne pas déséquilibrer la description d'un thème, qu'il soit développé dans une seule unité de base ou dans un regroupement de plusieurs unités. La décision de lier ou non deux segments non consécutifs suit le même principe que pour la segmentation ; le seuil de coupure est donc le même.

5. Annotation sémantique

Une fois les segments délimités et organisés entre eux, il s'agit d'informer l'utilisateur du contenu de chacun : quels sont ses thèmes, est-on dans une partie introductive, conclusive, etc. ? Ce type d'information sera donné par la sélection des termes les plus saillants ainsi que par la mise en évidence des phrases caractéristiques.

5.1. Les annotations discursives

L'annotation discursive est réalisée en appliquant les connaissances linguistiques décrites précédemment. Plus précisément, les règles d'exploration contextuelle ont pour fonction d'attribuer une étiquette aux phrases qui introduisent des cadres thématiques ou organisationnels. Après le déclenchement de l'ensemble des règles, la représentation du texte construite par l'analyseur (cf. § 6) est ainsi décorée par ces étiquettes. Ce sont ces décorations, entre autres, qui vont être exploitées par le module de visualisation.

5.2. Extraction des termes saillants

Plusieurs travaux de recherche se sont intéressés aux moyens de décrire un texte ou une de ses portions, certains par classification (Ferret *et al.*, 1998), d'autres en extrayant les phrases les plus significatives (Mani *et al.*, 1999), d'autres encore en identifiant les thèmes significatifs.

La majorité des auteurs travaillant sur l'identification de thèmes s'appuie sur les groupes nominaux pour décrire des thèmes d'un texte (Mather *et al.*, 2000 ; Boguraev *et al.*, 1997 ; Lin *et al.*, 1997), se différenciant principalement dans leur manière de mesurer la proéminence thématique de ces expressions. Ainsi pour certains, la pertinence de ces expressions dépend de leur mise en valeur par des marques linguistiques (Porhiel, 2001), pour d'autres, de leur position relative et absolue, et ce en fonction du type de document (Lin *et al.*, 1997), et pour d'autres enfin, plus classiquement, de leur fréquence et de leur distribution (Boguraev *et al.*, 1997).

Les travaux de (Boguraev *et al.*, 1997 ; Boguraev *et al.*, 1999) sont les plus avancés dans le domaine. Ils proposent de présenter les segments de texte par les groupes nominaux représentatifs des entités les plus significatives du segment considéré, accompagnés de spécifications contextuelles de différents niveaux de granularité (groupe verbal, proposition minimale, phrase).

Notre système d'identification des termes saillants s'inspire de leurs travaux. Il s'articule en trois étapes successives. La première étape vise à repérer les expressions « référentes », c'est-à-dire les formes nominales ou pronominales faisant référence aux entités du texte. La seconde étape est un système de résolution d'anaphores qui d'une part, permet de distinguer les formes descriptives des entités de leurs reprises anaphoriques et d'autre part, de gagner en précision sur la fréquence et la distribution des entités du texte. La troisième étape intervient pour caractériser la pertinence thématique de ces entités vis-à-vis de la structure du texte.

5.2.1. Repérage des expressions référentes

Le processus de repérage d'expressions référentes consiste à extraire des syntagmes nominaux et des pronoms par application de patrons syntaxiques. On distingue les syntagmes nominaux simples, seulement constitués de noms et d'adjectifs, des syntagmes nominaux complexes, admettant la préposition « de ».

5.2.2. Système de résolution d'anaphores

Nous avons ciblé nos traitements sur le repérage des expressions les plus susceptibles de jouer un rôle de reprise anaphorique, c'est-à-dire essentiellement les expressions nominales introduites par un démonstratif et les pronoms personnels dont nous avons filtré les tournures impersonnelles les plus courantes. Seuls des syntagmes nominaux sont envisagés comme antécédent potentiel.

Le système suit une architecture pipeline en trois phases : une sélection des antécédents candidats pour une anaphore donnée, un ordonnancement préférentiel des candidats pour désigner le plus probable et un filtrage des antécédents incompatibles avec l'expression référente en focus. Ces deux dernières étapes sont inversées par rapport à ce que l'on peut trouver dans la littérature (Mitkov, 1998 ; Boguraev, 1997). Ceci découle du fait qu'en grande majorité il existe un antécédent pour le type d'anaphore que nous traitons, et par là nous voulons forcer leur repérage malgré les contraintes de filtrage. Celles-ci sont donc ordonnées entre elles par degré de contraintes. Le premier antécédent qui valide le plus haut degré est considéré comme l'antécédent de l'anaphore.

Les critères de sélection, de filtrage et de préférence sont fondés sur des critères morphologiques (compatibilité genre/nombre), lexicaux (même tête sémantique), syntaxiques (fonction grammaticale, positions parallèles) et discursifs (récence) qui permettent de pondérer préférentiellement tel ou tel candidat antécédent. La phase de préférence calcule deux types de préférence : un poids inhérent au caractère propre de l'antécédent et un poids fonction du type de l'anaphore.

La plupart des ressources nécessaires pour réaliser ces descriptions d'expressions référentes ont été obtenues à partir d'heuristiques robustes : par exemple, l'annotation de la fonction grammaticale Sujet ou Objet est détectée selon la position de la forme par rapport au verbe de la proposition et selon sa position par rapport à une préposition.

5.2.3 Caractérisation des termes saillants

Parmi les entités obtenues, seules les entités présentant un « caractère thématique » sont retenues. Ce caractère se présente sous la forme d'un seuil minimal sur le poids d'une entité. Il a été fixé pour privilégier les fonctions grammaticales sujet et objet, la récurrence ou la spécification par un démonstratif comme caractères thématiques. Les caractéristiques d'une entité sont celles de ses individus moyennées.

Les entités thématiques sont ensuite caractérisées en fonction de leur rôle dans la structure thématique du texte. Nous utilisons le nombre de segments thématiques dans lesquels un terme apparaît pour qualifier le descripteur de local s'il apparaît dans un seul segment ou de global s'il figure dans au moins deux segments. L'objectif sous-jacent de ce typage vise à distinguer, au sein d'un même segment, les thèmes qui permettent de situer le segment dans l'ensemble du document des thèmes qui décrivent la spécificité du segment.

6. La plate-forme ContextO

L'annotation discursive est réalisée en appliquant les connaissances linguistiques décrites précédemment. Ces connaissances sont représentées dans le formalisme de l'exploration contextuelle (Desclés *et al.,* 1997 ; Minel *et al.,* 2001) et implémentées dans la plate-forme ContextO. Rappelons brièvement les caractéristiques de cette plate-forme qui se compose d'un gestionnaire de connaissances linguistiques, d'un analyseur de textes, d'un étiqueteur sémantique et d'agents spécialisés[5].

- Gestionnaire de connaissances linguistiques

Le système de gestion des connaissances linguistiques (Crispino *et al.*, 1999) a pour charge d'accueillir les connaissances linguistiques dans un modèle conforme à la méthode d'exploration contextuelle. Il est doté d'un interpréteur du langage de description de ces connaissances linguistiques, modèle fondé sur la notion d'indicateurs, d'indices et de règles d'exploration contextuelle.

- Analyseur de texte

L'analyseur de textes construit une représentation qui reflète l'organisation hiérarchique du texte, le traitement textuel nécessitant en effet qu'une tâche spécialisée puisse se focaliser sur les *n* unités lexicales de la *i*ème phrase du *j*ième

[5] Le lecteur trouvera une description détaillée de la plate-forme ContextO dans (Ben Hazez 2002 ; Minel 2002 ; Crispino 2003).

paragraphe de la *k*ième section. La construction de cette structure hiérarchisée s'appuie sur le texte segmenté produit par un segmenteur (Mourad, 1999) et un « tokeniseur » (Crispino *et al.*, 1999). L'analyseur ne vérifie pas la cohérence de cette structure puisque celle-ci a été produite par le segmenteur ou par un analyseur XML conformément à une DTD (Définition de Type de Document).

- Etiqueteur sémantique

Pour chacune des phrases qui composent le texte analysé, l'étiqueteur sémantique adresse une requête au gestionnaire des connaissances linguistiques afin que celui-ci identifie les positions des indicateurs et des indices et le nom des règles d'exploration contextuelle qui peuvent être déclenchées. L'étiqueteur déclenche l'exécution des règles, chaque règle étant encodée comme une méthode d'une classe Java. L'étiqueteur sémantique déclenche, pour toutes les tâches choisies par l'utilisateur, toutes les règles associées à celles-ci. Les règles sont en effet considérées comme indépendantes et l'ordre de leur déclenchement, pour une tâche donnée, est donc indifférent. Ce mode de fonctionnement correspond à l'hypothèse que, pour une tâche donnée, certains marqueurs sémantiques ne sont pas exclusifs entre eux.

- Agents spécialisés

Les agents spécialisés sont munis de connaissances et de modèles de présentation des résultats de l'étiquetage qui sont dédiés à des tâches spécifiques. L'agent de visualisation présenté dans la section 7 est une illustration de cette notion.

La coopération de ces quatre sous-systèmes s'articule autour d'un modèle du texte (Crispino *et al.,* 1999). C'est ce modèle du texte qui va être décoré avec les résultats produits par le moteur d'exploration contextuelle en appliquant les règles d'exploration contextuelle. Ce modèle est construit à partir du traitement automatique des balises qui sont présentes dans le texte ou qui seront dynamiquement construites en analysant les marques de surface du texte. Il ne peut plus ensuite être altéré dans sa structure par l'application des règles d'exploration conceptuelle. Ce choix conceptuel est motivé par le souci d'offrir un noyau commun et stable aux différents systèmes, éventuellement externes à ContextO, qui coopèrent. Soulignons que ce choix ne signifie pas qu'il ne soit pas possible de créer d'autres structures qui viendraient se greffer sur ce modèle, mais simplement que ces nouvelles structures ne peuvent être exploitées que par d'autres tâches.

7. Visualisation

Notre conception des outils de navigation, dans le cadre de la fouille de textes, repose sur un principe général concernant l'objet textuel à produire : à partir de l'objet textuel source T, le résultat de la fouille de texte est un nouvel objet textuel T_f qui intègre l'objet T et une représentation décorée de T appelée T_d. Par conséquent, un déplacement par l'utilisateur dans T_d doit toujours être répercuté par le système

de navigation dans T, ce qui implique une représentation du texte et un langage associé.

Notre représentation d'un texte (Crispino *et al.*, 1999 ; Crispino 2003) cherche à représenter la structure logique de la manière la plus générale possible en incluant les éléments textuels qui peuvent être le mieux exploités (Minel *et al.,* 2001). Notre vision d'un texte est donc celle d'une hiérarchie simple d'éléments textuels. Bien qu'il soit fréquent de trouver des textes qui présentent une organisation plus complexe, nous pensons que notre option fournit, d'une part un cadre suffisant pour la formulation des connaissances linguistiques exprimables dans notre méthode et d'autre part, constitue un modèle de base facilement extensible pour traiter des phénomènes plus complexes. Nous modélisons cette hiérarchie par une structure d'arbre. La racine représente le texte complet. Les descendants directs de la racine représentent les sections de premier niveau dans le texte. C'est-à-dire qu'en parcourant de gauche à droite les descendants directs de la racine, nous nous positionnons de manière ordonnée dans les sections de premier niveau du texte. Les sections peuvent avoir comme descendants d'autres sections de niveau plus bas (sous-sections) ou des paragraphes de la section. Les descendants des paragraphes sont les nœuds qui représentent les phrases et finalement, les descendants des phrases, les feuilles de l'arbre, sont les unités lexicales.

Pour manipuler cette structure un langage a été défini. Il s'agit d'un langage formel de type déclaratif, structuré en plusieurs couches : une couche de niveau inférieur, le noyau, assurant les fonctionnalités de base et plusieurs couches de plus haut niveau assurant les fonctionnalités propres à des tâches spécifiques. Le lecteur pourra se reporter à (Crispino, 2003) pour une présentation plus approfondie.

Du point de vue des interfaces visuelles, plusieurs objectifs (Couto, 2002) ont été visés. En premier lieu, l'utilisateur doit pouvoir fouiller plusieurs textes. En second lieu, chaque document doit pouvoir être visualisé selon des vues graphiques différentes : une vue linéaire et des vues reflétant l'organisation du texte, que ce soit l'organisation logique intrinsèque au texte ou la structure thématique. En troisième lieu, pour un même document, l'utilisateur doit pouvoir définir différents profils de fouilles. Un profil correspond au choix de différentes décorations pour une tâche donnée (Berri *et al.*, 1996 ; Minel *et al.*, 2001). Enfin, pour chaque document l'utilisateur peut demander des vues graphiques spécialisées qui sont autant de nouvelles formes sémiotiques du texte T_d. D'une manière générale, la flexibilité et la dynamique dans l'affichage des différentes vues constituent des objectifs d'ergonomie générale.

Notre modèle général de l'écran de navigation repose sur les principes fondamentaux de toute interface utilisateur (GUI) (Shneiderman 1998 ; Hearst 1999), que nous supposerons connus.

L'écran de navigation (cf. figure 5, qui visualise le texte pris comme exemple dans la section suivante) se compose, en plus des barres standard, de quatre fenêtres, A, B, C, D. La fenêtre A visualise la liste des documents choisis par l'utilisateur.

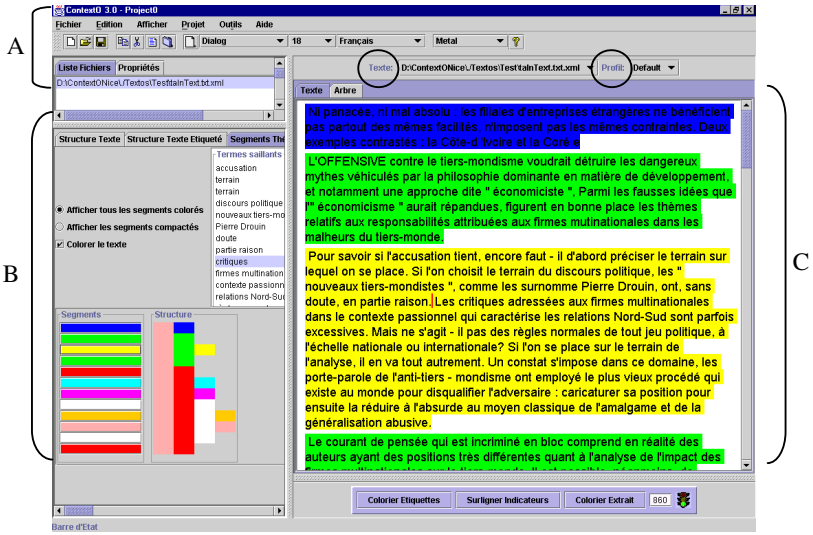

Figure 5. *Exemple d'interface dynamique*

La fenêtre B offre différentes vues décorées du texte source T, accessibles par des onglets spécifiques. Ces décorations sont le résultat des calcul effectués (repérage des segments thématiques, fréquence des termes saillants, étiquette discursive attribuée à une phrase, etc.) et offre des points d'accès aux différentes parties du texte. Chaque segment thématique identifié est visualisé sous la forme d'un rectangle coloré, les segments liés à un même thème étant de la même couleur. Sa taille en nombre de phrases est disponible afin d'évaluer ce segment ou ce thème par rapport aux autres thèmes du texte. La visualisation de la structure permet de mieux indiquer l'organisation des segments entre eux. L'inclusion d'un segment dans un autre signifie qu'il y a un lien thématique entre eux et permet de visualiser ainsi sa place par rapport aux autres thèmes. Pour chaque segment, il est possible de visualiser la liste des termes saillants avec leur fréquence. D'autres onglets permettent d'obtenir une vue hiérarchique du texte, organisé en arbre, ou une vue globale des annotations attribuées aux phrases. Cette fenêtre offre donc différents résumés du texte : selon les thèmes, selon les marques discursives, ou selon sa structure logique. Il est ainsi possible de prendre connaissance globalement du contenu du texte sans avoir à le lire et d'approfondir seulement certains points si des parties semblent intéressantes.

La fenêtre C visualise le texte source en cours de fouille et la fenêtre D montre le fragment textuel construit à partir de l'application d'une tâche et d'un profil de fouille choisis par l'utilisateur, qui n'est pas illustré sur la figure 5 pour des raisons

de lisibilité. Toutes ces fenêtres sont liées de telle manière qu'une modification dans une des fenêtres se propage dans les autres.

Les fenêtres C et D présentent une interaction plus spécifique que nous allons décrire. Le texte source est visualisé dans la fenêtre supérieure et le fragment textuel dans la fenêtre inférieure. Des options de visualisation (les deux fenêtres intitulées *Texte* et *Profil* sur la figure 5) sont offertes pour permettre à l'utilisateur de visualiser plusieurs documents avec pour chaque document, des profils différents, suivant le principe du "mille-feuilles". La dynamique entre les deux fenêtres permet ainsi à l'utilisateur de visualiser constamment le contexte d'une phrase du fragment textuel, indépendamment du type de la vue affichée (linéaire ou structurelle). D'autre part, la visualisation des informations saillantes repose sur des principes d'ergonomie qui utilisent les possibilités offertes par la colorisation (Murch, 1985). La saillance et les décorations sémantiques sont traduites par des couleurs choisies par l'utilisateur. Un menu contextuel permet d'afficher la ou les étiquettes attribuées et de se déplacer dans le texte en recherchant les phrases annotées identiquement en précisant la portée relative (*dans un paragraphe, dans une section, dans le texte*) ou absolue (*première ou dernière phrase du texte*) de la recherche. Nous travaillons actuellement à un langage qui possède des fonctions de navigation prennant en compte des organisations textuelles plus complexes.

Toutes ces fonctionnalités ont été développées dans un composant autonome en Java, en s'appuyant sur les principes de conception d'un Bean[6]. Ce composant est utilisé dans la plate-forme ContextO.

Les principes de navigation et de visualisation que nous avons décrits mettent en avant le principe de l'interaction plutôt que celui d'une focalisation par granularité de plus en plus fine comme dans (Boguraev *et al.*, 2001). Nous pensons en effet que l'activité consistant à sélectionner des informations à l'intérieur d'un texte donné est une démarche hautement intelligente qui varie en fonction des sujets, de leur expertise du domaine traité ainsi que de leur degré d'attention au moment où ils en prennent connaissance. Actuellement cette activité n'est pas modélisable dans sa globalité. En revanche, nous pouvons viser la construction de représentations d'un texte qui fournissent à l'utilisateur des repères sur les organisations discursives qui structurent le texte. Ce repérage permettra à l'utilisateur de déployer ses propres stratégies de recherche d'informations.

8. Résultats

Nous allons ici présenter les résultats de notre système sur un texte. Notre choix s'est porté sur un texte de quatre pages du *Monde Diplomatique* (voir figure 6) qui développe le rôle des firmes internationales dans la lutte contre le tiers-mondisme. Ce texte possède un titre, mais aucun sous-titre. Nous avons conservé les marques de

[6] Néanmoins ce composant ne constitue pas un Java Bean au sens strict.

paragraphes, qui constituent les unités de base de la segmentation (nous ne présentons ici que l'extrait[7] le plus illustratif de la collaboration des deux approches).

Les lexies en gras sont des introducteurs thématiques, ils sont précédés de *<intrThem i/>*. *<segCoLex i>* et *</segCoLex i>* sont les marques après segmentation thématique. **<Si>** et **</Si>** encadrent les segments définitifs.
En italique, les thèmes ayant une portée globale à plus d'un segment du texte, ceux qui ont une portée globale à un bloc sont mentionnés figure 7.
<segCoLex 8> => **<S8>**

[§11] L'importance quantitative de *l'investissement* étranger est cependant moins significative de l'impact des firmes multinationales que le type de secteurs où elles se localisent. En Côte-d'Ivoire, les firmes contrôlent pratiquement l'ensemble de l'industrie produisant pour le marché interne. Au contraire, l'accès à ce dernier leur est interdit dans la plupart des branches en Corée du Sud. Cette situation a des conséquences décisives, particulièrement sur trois variables stratégiques du processus de développement : *l'allocation des ressources*, le *modèle de consommation* et *l'intégration en amont de l'activité industrielle*.

<intrThem 1/> => **</S8> <S9>**

[§12] **En ce qui concerne** *l'allocation des ressources*, en Côte-d'Ivoire, l'excédent prélevé par l'Etat et consacré à l'expansion du marché intérieur transite nécessairement par les firmes multinationales, finançant en grande partie leur *implantation* ou l'élargissement de leur capacité productive. En Corée du Sud, les firmes multinationales sont exclusivement concentrées dans les branches exportatrices, ce qui permet à l'Etat de prélever des ressources externes additionnelles que les entreprises publiques ou privées coréennes utilisent selon les orientations précises du plan dans le cadre d'une stratégie d'intégration industrielle orientée vers le marché intérieur.

</segCoLex 8>
<segCoLex 9> => **</S9> <S10>**
<intrThem 2/>

[§13] **QUANT au** *modèle de consommation*, en Côte-d'Ivoire, la production de biens relève de la stratégie propre à la firme multinationale, sans rapport avec le niveau moyen des revenus et les habitudes traditionnelles de consommation. Ce phénomène suscite ou accentue à son tour la *distribution* inégalitaire du revenu. En Corée du Sud, la diversification des biens offerts aux *consommateurs* est un processus progressif et contrôlé en relation étroite avec la capacité d'achat de la population. Cette correspondance entre niveau de revenu et offre de biens contribue fortement à atténuer les tendances à la répartition inégalitaire des revenus. La politique du pouvoir, dans ce domaine, a été très ferme. Les biens de consommation les plus modernes - *électroménager*, appareils optiques, électronique grand public, - fabriqués en grande partie par les firmes multinationales, ont été longtemps exclusivement destinés à *l'exportation*. La population coréenne n'a eu accès à ces *biens* qu'une fois satisfaits les besoins essentiels en matière de nourriture et de vêtement. Mais le développement du marché interne n'a pas profité aux firmes multinationales qui en ont été pratiquement exclues au profit des firmes locales. Dans la branche électronique grand public, par exemple, les ventes sur le marché interne sont réalisés pour 95,4 % par les entreprises coréennes et pour 4,4 % par des entreprises en joint venture.

<intrThem 3/> => **</S10> <S11>**

[§14] Enfin, **pour ce qui concerne** *l'intégration en amont de l'activité industrielle*, dans un pays comme la Côte-d'Ivoire, où le secteur industriel est contrôlé par les firmes étrangères, la taille du marché a constitué l'obstacle insurmontable à la diversification de la structure productive. En conséquence, le processus reste bloqué au niveau des branches légères. En Corée du Sud, la maîtrise absolue de l'Etat sur

[7] Le texte complet peut être consulté à
http://www.limsi.fr/Individu/hernandz/local/Download/lesFirmesMultinationales.txt

la décision économique au niveau du marché interne a permis ce que l'on appelle la "remontée des filières" vers les industries lourdes - sidérurgie, chimie et industries de biens d'équipement - et assuré une autonomie notable du processus d'industrialisation, même si, dans certains secteurs, la dimension du marché était manifestement insuffisante.

</segCoLex 9>
<segCoLex 10> **=> </S11> <S12>**

[§15] Cette rapide comparaison montre que le diagnostic établi par les *analystes des problèmes du développement* n'est pas aussi faux que cela et que la thérapie proposée, qui est une "thérapie douce", loin de conduire à des situations apocalyptiques, peut se révéler efficace.

Figure 6. *Extrait d'un texte du Monde Diplomatique*

Si l'on se ramène à la figure 2 de la section 4.1, l'application du critère de cohésion lexicale conduit à la production de 10 segments, où les paragraphes 4 et 5, 6 et 7, 9 et 10, 11 et 12 et 13 et 14 sont regroupés. Après recherche des marques indicatrices d'un début de thème, seuls les introducteurs débutant les paragraphes 12, 13 et 14 sont trouvés. Quand on aligne les 2 résultats, les marques *<intrThem 1/>* et *<intrThem 3/>* sont toutes deux incluses dans un segment, ce qui provoque des divisions en 2 segments différents. La marque *<intrThem 2/>* coïncide quant à elle avec un début de segment. La rectification effectuée est tout à fait cohérente par rapport au texte. En effet, ces 3 paragraphes abordent des thématiques différentes, et si ils sont reliés en deux occasions, c'est essentiellement dû à la présence des mots côte d'Ivoire et Corée qui se répètent souvent à cet endroit du texte, en liaison avec la notion de marché interne (§13 et §14) ou en liaison avec l'introduction des 3 thématiques qui entraîne le lien (§11 et §12). En lisant le texte, on s'aperçoit que les autres décisions de segmentation sont tout à fait justifiées. On obtient donc, après intégration, 12 segments que l'on peut voir sur la figure 5 de la section précédente.

Le module de structuration construit ensuite les emboîtements illustrés par la figure 7. S1, en tant qu'introduction, reste isolé puisque les autres segments vont développer le propos général du texte. Le premier bloc délimité correspond à une analyse générale du développement des multinationales. Le bloc suivant correspond à l'annonce du type de rôle que peuvent avoir les firmes avec S5, relié à la conclusion en S12, alors que S6 annonce l'étude des deux exemples, la Corée et la Côte d'Ivoire, et S7, des considérations générales sur ces deux pays. Le bloc suivant correspond à l'étude des différents points. S8 annonçant l'analyse faite au niveau des marchés internes et des trois points étudiés en S9, S10 et S11, il est logique que l'ensemble soit englobé dans une même thématique.

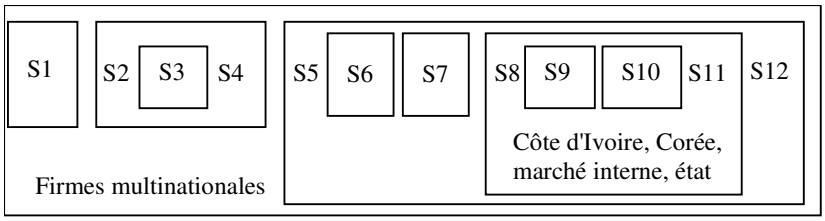

Figure 7. *Structure produite*

Parmi les termes saillants, il existe des termes répartis sur des blocs (cf. figure 7), et des termes repris dans différents segments (cf. termes en italique figure 6). Au titre des caractéristiques de ce texte, les différents thèmes sont d'abord présentés de manière générale en S4 et repris dans la suite du texte dans les exemples, et que l'on retrouve donc dans les segments S8 à S11. En ce qui concerne les types de termes, on retrouve les notions abordées, avec des termes tels que *investissement, problème de développement, modèle de consommation*, mais aussi des termes ayant davantage un rôle d'organisation et de caractérisation du discours, avec des termes tels que *analyse* et *constat*.

D'un point de vue plus général, nous avons appliqué notre méthode sur des textes différents, issus des revues *la Recherche* et *Pour la Science*, avec des résultats tout à fait cohérents. La segmentation thématique a fait par ailleurs l'objet d'évaluations propres (Ferret, 2002). Enfin, notre méthode de segmentation-structuration permet bien de retrouver une organisation globale hiérarchique du texte lorsque celle-ci est présente. Le critère de cohésion lexicale semble approprié pour ce type de tâche, ce qui s'explique par le fait que la reprise d'un thème, pour conclure ou pour annoncer un nouveau développement, s'effectue en général par la répétition des termes qui le caractérise, surtout lorsque cette reprise est effectuée à une certaine distance dans le texte.

Afin d'évaluer plus précisément nos résultats, nous travaillons actuellement à l'élaboration d'un protocole permettant d'évaluer la présentation des textes par rapport à une tâche utilisateur du type recherche d'information.

9. Conclusion

Nous avons présenté dans cet article l'ensemble des processus nécessaires à une présentation du contenu d'un texte qui soit à la fois informative et indicative et qui permette une navigation interne au texte. Nous avons voulu donner à un utilisateur la possibilité de choisir le type d'exploration qui convient le mieux à la tâche qu'il accomplit. Il peut ainsi visualiser l'ensemble des thèmes traités dans le texte, leur organisation, naviguer d'un passage à l'autre, ces passages pouvant être soit des

phrases jugées importantes ou bien des segments thématiquement homogènes. Les informations nécessaires à ces traitements sont automatiquement calculées et se concrétisent par l'ajout d'annotations sémantiques dans le texte. Elles proviennent des résultats d'une analyse thématique faisant collaborer deux processus. L'un se fonde sur la cohésion lexicale et l'autre sur le repérage de marques linguistiques. Une autre originalité de notre approche repose sur la construction automatique de la structure du texte, qui est indépendante de l'existence d'une structure logique préalable. Des annotations étiquetant le rôle discursif de certaines phrases ou les termes saillants des segments ou de blocs de segments sont aussi rajoutées automatiquement dans le texte.

L'ensemble de ces travaux a été implémenté dans la plate-forme ContextO et a permis la réalisation d'une première version d'une interface adaptative. L'utilisateur peut effectivement passer d'une entité à une autre, revenir au texte et se définir un profil de fouille de texte. Il a ainsi la possibilité de parcourir le texte rapidement afin de prendre connaissance des thèmes développés, en utilisant les résultats de l'analyse thématique. Si certains segments l'intéressent plus particulièrement, il peut y accéder soit intégralement, soit en faisant afficher seulement certaines de leurs phrases caractéristiques.

Les perspectives que nous envisageons concernent l'expérimentation de la plate-forme par des utilisateurs dans le but d'évaluer l'interface et les informations dispensées, à la fois du point de vue de leur qualité mais aussi de leur utilité pour trouver l'information cherchée. Le principe serait de définir des tâches précises telles que répondre à une question factuelle, rechercher les résultats énoncés dans un texte ou bien reconnaître les thèmes principaux du texte, ces tâches permettant d'évaluer des utilisations différentes de la plate-forme.

Remerciements

Nous remercions tous les participants du projet RÉGAL, et plus particulièrement M. Charolles, pour leurs contributions au projet. Nos remerciements s'adressent aussi à F. Prévitali et A. Tessier qui ont participé aux développements des interfaces dédiées au projet dans le cadre de leur stage respectif de l'IIE et du DESS ILSI de l'Université Paris-Sorbonne. La plate-forme ContextO, est le fruit de la collaboration entre l'Université Paris-Sorbonne et l'Université de la République (Uruguay) et a reçu le soutien du programme Ecos-Sud. Le projet RÉGAL a reçu le soutien de l'ACI Cognitique (LAC038).

10. Bibliographie

Asher N., and Lascarides A., *Intentions and information in discourse*, Proceedings of the 32nd ACL, pp. 34-41, 1994.

Ben Hazez S., *Un modèle d'exploration contextuelle des textes : filtrage et structuration d'informations textuelles, modélisation et réalisation informatique (système SEMANTEXT)*, Thèse de doctorat, Université Paris-Sorbonne, Paris, 2002.

Berri J., Cartier E., Desclés J.-P., Jackiewicz A., Minel J.-L., « Safir, système automatique de filtrage de textes », *Langages, Cognition et Texte* , vol II., Université Hankuk des Langues étrangères, Université Paris-Sorbonne, Séoul et Paris, pp. 3-16, 1996.

Boguraev B. C., Kennedy R., Bellamy R.K., Brawer S., Wong Y., Swart J., « Dynamic Presentation of Document content for Rapid On-Line Skimming », *AAAI Spring Symposium on Intelligent Text Summarisation*, Stanford, CA, 2001.

Boguraev B., Bellamy R. K., Kennedy E C., « Dynamic Presentation of Phrasally-based Document Abstractions », *Proceedings of the 32nd HICSS*, 1999.

Boguraev B., Kennedy C., « Salience-based content characterisation of text documents », In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, Madrid, Spain, pp. 2-9, 1997.

Charolles M., « L'encadrement du discours - Univers, champs, domaines et espaces », *Cahier de recherche linguistique*, 6, LANDISCO, URA-CNRS 1035 Université Nancy 2, 1997.

Chen H., Houston A., Sewell R., Schatz B., Internet browsing and searching: user evaluations of category map and concept space techniques, *Journal of the American Society for Information Science*, Special Issue on "*AI Techniques for Emerging Information Systems Applications*" 49 (7), pp. 582-603, 1998.

Choi F., « Advances in domain independent linear text segmentation », *Actes de NAACL'00*, pp. 26-33, 2000a.

Choi F., « A speech interface for rapid reading », *Proceedings of IEE colloquium: Speech and Language Processing for Disabled and Elderly People*, London, England, 2000b.

Couto J., *ContextO, Los sistemas de exploracion contextual de cara al usuario*, Mémoire de Master, Université de la République, Uruguay, 2001.

Crispino G., Desclés J.-P., Ben Hazez S., Minel J.-L., « Architecture logicielle de Context, plate-forme d'ingénierie linguistique », *TALN 99,* Cargèse, pp. 327-332, 1999.

Crispino G., *Conception et réalisation d'un système informatique d'exploration contextuelle. Conception d'un langage de spécification de connaissances linguistiques*, Thèse de doctorat, Université Paris-Sorbonne, Paris, 2003.

Cutting D., Karger D.,. Pedersen J, Tukey J., « Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections », *Proceedings of the 15th Annual International ACM/SIGIR Conference*, Copenhagen, 1992.

Desclés J.-P., Cartier E., Jackiewicz A., Minel J.-L., « Textual Processing and Contextual Exploration Method » in *CONTEXT 97*, Universidade Federal do Rio de Janeiro, Brésil, pp. 189-197, 1997.

Ferret O., « Using collocations for topic segmentation and link detection », *Actes de COLING 2002*, pp. 260-266, 2002.

Ferret O., Grau B., Minel J.-L. , Porhiel S. , « Repérage de structures thématiques dans des textes », *TALN* 2001, Tours, pp. 163-172, 2001.

Ferret O., Grau B., « A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts », *ECAI*, pp. 155-159, 1998.

Ferret O., Grau B., Masson N., « Thematic segmentation of texts: two methods for two kinds of texts », *Actes de ACL-COLING'98*, pp. 392-396, 1998.

Fløttum K., Quant à – *Thématiseur et focalisateur*, Actes du colloque de Caen 9-11 octobre 1997, La thématisation dans les langues, Berne, Peter Lang, pp. 135-159, 1999.

Grosz B., Sidner C., *Attention, Intentions and the structure of discourse*, Computational Linguistics, 12(3), pp. 175-204, 1986.

Halliday M.A.K., Hasan R., *Cohesion in English*, London, Longman, 1976.

Hearst M., *User Interfaces and Visualization*. In Modern Information Retrieval R. Baeta-Yates, B. Ribeiro-Neto (eds), Addison-Wesley, pp. 257-322, 1999.

Hearst M., TextTiling: *Segmenting Text into Multi-paragraph Subtopic Passages*, Computational Linguistics, vol. 23, n° 1, pp. 33-64, 1997.

Hearst M. A., *Improving Full-Text Precision on Short Queries using Simple Constraints*. In Proceedings of the Symposium on Document Analysis and Information Retrieval, April 1996.

Hernandez N., Grau B., *Analyse thématique du discours : segmentation, structuration, description et representation*, Actes de CIDE'2002, Hammamet, Tunisie, pp. 277-285

Hobbs J.R., Stickel M., Appelt D. and Martin P., *Interpretation as abduction*, Artificial Intelligence, vol. 63, pp. 69-142, 1993.

Illouz G., *Vers un apprentissage en TALN dépendant du type de texte*, TALN 2000 (Traitement Automatique du Language), Lausanne, Suisse, 2000.

Jackiewicz A. *L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle*. Thèse de Doctorat, Université Paris-Sorbonne, 1998.

Jackiewicz A., Minel J.L., *L'identification des structures discursives engendrées par les cadres organisationnels*, TALN 2003, Batz-Sur-Mer, 2003.

Jackiewicz A. *Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes*, CIFT'02*, pp.* 95-107., Hammamet*,* Tunisie, 2002.

Jacquemin C., Jardino M., *Multi-dimensional and Multi-scale Visualizer of Large XML Documents*, Proceedings of EUROGRAPHICS*,* Saarbrücken, Germany, 2002.

Kan M-Y., McKeown K., Klavans J., *Domain-specific informative and indicative summarization for information retrieval*, Proceedings of the first Document Understanding Conference (DUC), pp.19-26 , New Orleans, 2001.

Kozima H., Text Segmentation Based on Similarity between Words », Actes de ACL'93, pp. 286–288, 1993.

Le Priol, F. *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK -JAVA : identification et interprétation de relations entre concepts,* Thèse de Doctorat, Université Paris-Sorbonne, 2000.

Lin C-Y., Hovy E., *Identify Topics by Position*, Proceedings of the 5th Conference on Applied Natural Language Processing, March 1997.

Mani I., Maybury M. T., *Advances in automatic text summarization,* MIT Press, Cambridge, MA, 1999.

Mann W. C., Thompson S. A., *Rhetorical Structure theory: toward a functional theory of text organization*, Text, 8(3), pp. 243-281, 1988.

Marcu D. *The rhetorical parsing, summarization, and generation of natural language texts*, Ph.D. Dissertation, Department of Computer Science, Universtity of Toronto, 1997.

Marcu D., *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, November 2000.

Mather L. A., Note J., *Discovering Encyclopedic Structure and Topics in Text*, Sixth ACM SIGKDD. Boston, MA, USA, August 2000.

Minel J.-L. *Filtrage sémantique de textes.* Hermès, 2003.

Minel J.-L. *Filtrage sémantique de textes. Problèmes conception et réalisation d'une plate-forme informatique.* Habilitation à diriger des recherches, Université Paris-Sorbonne, 2002.

Minel J.-L., Desclés J-P., Cartier E., Crispino G., Ben Hazez S., Jackiewicz A., *Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText*, Revue Technique et Science informatiques, Hermès, n° 3, pp. 369-395, 2001.

Mitkov R., *Robust pronoun resolution with limited knowledge*. In Proceedings of the 18th International Conference on Computational Linguistics, COLING/ACL, 1998.

Mourad G. *La segmentation des textes par l'étude de la ponctuation*, CIDE'99, Damas, Syrie, 1999.

Murch G. J. *Color Graphics - Blessing or Ballyho*, in Baecker *et al.* (eds) Readings in Humanities Computer Interaction : Toward the Year 2000. Morgan Kaufman Publishers, 1985.

Passonneau, J. R., Litman, J. D., *Discourse Segmentation by Human and Automated Means*, Computational Linguistics, vol. 23, n° 1, *ACL*, 1997.

Porhiel S. *Les introducteurs de cadre thématique*, Cahiers de Lexicologie 83, 2 : 1-36, 2003.

Porhiel S. *Linguistic expressions as a tool to extract thematic information*, Actes Corpus Linguistic, Lancaster University, 2001.

Saggion H., et Lapalme G., *Concept Identification and Presentation in the Context of Technical Text Summarization*, RIAO, Paris, France, 2000.

Salton G., Singhal A., Buckley C. and Mitra M., *Automatic Text Decomposition Using Text Segments and Text Themes*, Actes Hypertext'96, Seventh ACM Conference on Hypertext, Washington, D.C., pp. 53-65, 1996.

Shneiderman B. *Designing the User Interface : strategies for effective Human-Computer interaction*. Addison-Wesley, 1998.

Wise J. A., Thomas J. J., Pennock K., Lantrip D., Pottier M., Schur A., and Crow V., *Visualizing the non-visual: Spatial analysis and interaction with information from text documents*. Proc. of Information Visualization. IEEE Computer Society Press, 1995.

Yaari Y. *Segmentation of Expository Texts by Hierarchical Agglomerative Clustering*, Proceedings of RANLP ,1997.