La segmentation thématique TextTiling comme indice pour le repérage de segments d'information évolutive dans un corpus de textes encyclopédiques

Marion LAIGNELET[1,2], Christophe PIMM[3]
[1] CLLE-ERSS – Université Toulouse 2 – Le Mirail, Toulouse
[2] Société INITIALES – Montpellier
[3] CLLE-ERSS – Université Toulouse 2 – Le Mirail, Toulouse
{marion.laignelet,christophe.pimm}@univ-tlse2.fr

Résumé. Nous faisons l'hypothèse que les bornes délimitées par la méthode statistique TextTiling peuvent servir d'indices qui, cumulées à des indices de nature linguistique, permettront de repérer automatiquement des segments d'informations évolutives. Ce travail est développé dans le cadre d'un projet industriel plus général dont le but est le repérage automatique de zones textuelles contenant de l'information potentiellement évolutive.

Abstract. Our hypothesis is that the TextTiling's boundaries can be considered as clues we can use with other linguistic features to automatically detect evolving information segments. This work is developed as part of an industrial project aiming to automatically detect textual zones containing potentially evolving information.

Mots-clés : segments d'information évolutive, segmentation, algorithme TextTiling.

Keywords: evolving information, segmentation, TextTiling algorithm.

1 Introduction

Un segment d'information évolutive (SEDIS-ε) est une portion de texte de longueur variable contenant de l'information dont la particularité est d'être susceptible d'évoluer dans le temps. Cette notion d'évolutivité de l'information a pris naissance dans un contexte industriel particulier, l'édition, et plus précisément dans le cadre d'une problématique de mise à jour éditoriale de documents encyclopédiques. L'étude présentée dans cet article est partie prenante d'un projet plus global[1] dont le but applicatif final est de proposer à des rédacteurs chargés de mettre à jour les articles encyclopédiques un outil facilitant cette tâche de mise à jour de l'information. Cet article vise à montrer que l'association de méthodes linguistiques avec des méthodes statistiques peut représenter un apport non négligeable lorsqu'il s'agit de rendre compte de phénomènes complexes. Plus précisément, nous présentons une

[1] Projet mené dans le cadre d'une thèse CIFRE, partenariat entre l'ERSS, Toulouse, et la société Initiales, Montpellier.

expérimentation dans laquelle nous exploitons la notion de rupture thématique issue du TextTiling (Hearst, 1994) avec l'idée que cela peut contribuer à l'identification des segments recherchés (les SEDIS-ε). Nous proposons ainsi un point de vue particulier sur la notion de segmentation des textes en optant pour une vision à gros grain de la segmentation, puisque d'un côté nous cherchons à décrire les SEDIS-ε, et puisque de l'autre, le TextTiling a pour but de segmenter les textes en fonction des thèmes qui sont abordés dans leurs parties. Notre hypothèse est que les ruptures thématiques fournies par cet algorithme peuvent devenir des indices pour aider au repérage des frontières initiales et/ou finales de certains types de SEDIS-ε. La première partie de cet article présente la notion de segmentation et son importance lorsqu'on travaille en linguistique du discours. Dans la seconde partie, nous présentons la notion de SEDIS-ε ainsi que l'apport d'une méthode telle que le TextTiling pour la description et le repérage de tels segments de discours. La troisième partie fait état du protocole expérimental que nous avons suivis. Enfin, nous présentons les résultats et proposons une discussion dans la dernière partie.

2 Segmentation

Ce projet peut être (entre autres) situé à la fois dans le sillon de la Recherche d'Information (et de la recherche intradocumentaire, par extension) parce que nous avons comme objectif de rechercher parmi la masse de textes encyclopédiques ceux pour lesquels une mise à jour est nécessaire, et dans celui de l'Extraction d'Information puisque nous souhaitons rechercher et extraire une information qui précisément présente la caractéristique d'être évolutive. Une des techniques commune à ces deux domaines est la segmentation[2].

2.1 Point de vue linguistique et cognitif

Nous suivons Péry-Woodley (2005) lorsqu'elle écrit : *« Toute structuration passe en effet par une segmentation, segmenter impliquant à la fois diviser et regrouper en fonction d'un critère organisationnel. Que l'on envisage l'organisation discursive en termes de structure d'information, de structuration thématique ou de relations de cohérence, la notion de segmentation est présente : recherches de critères de regroupement d'unités (en segments), identification de marques de rupture ou de discontinuité (entre segments), étude des relations (entre segments) qui les hiérarchisent et forment des segments de niveau supérieur. L'identification de segments à même de présenter une homogénéité sémantique et/ou de constituer des unités fonctionnelles est également au coeur des recherches en T.A.L. touchant au discours. »* (Péry-Woodley, 2005)

Il est important de souligner que, sous le terme de « segmentation », nous ne nous limitons pas à la notion de « segmentation thématique », laquelle s'appuie sur la notion de cohésion lexicale (*cf.* Halliday et Hasan, 1976), c'est-à-dire la répétition des mots comme indicateur d'homogénéité thématique. De nombreux travaux, qu'ils soient menés en linguistique ou en psycholinguistique, cherchent à rendre compte de ces processus de construction de la cohérence que ce soit à travers la délimitation de segments de discours ou de la description des relations entre ces segments (*cf.* RST, SDRT, segmentation thématique, etc.). La segmentation automatique apparaît alors comme une technique incontournable développée relativement à divers objectifs applicatifs : pour certains l'objectif visé est celui du résumé automatique (*cf.* Minel (2002), Marcu (2000), etc.) ; pour d'autres il s'agit de développer des

[2] La segmentation peut également être vue comme une tâche à part entière, parti que nous n'adoptons pas dans ce cadre. Nous l'envisageons comme une étape nécessaire à un grand nombre de traitements sur les textes.

systèmes de recherche/sélection d'information importante (*cf.* Rossi et Bert-Erboul (1991), etc.) ou encore des outils d'aide à la navigation (*cf.* Jackiewick et Minel (2003), etc.). Dans tous ces cas, la tâche de segmentation est envisagée par rapport à une application précise, un point de vue sur les textes (*cf.* Hernandez, 2004). C'est également la position que nous adoptons : il nous paraît essentiel et incontournable de définir les segments que nous recherchons, les SEDIS-ε, ainsi que les méthodes et techniques mises en œuvre pour y aboutir relativement à l'objectif applicatif de départ, à savoir la mise en place d'un outil d'aide à la mise à jour de l'information dans des documents encyclopédiques.

2.2 Segmentation thématique et TextTiling

La segmentation thématique est un domaine riche pour lequel il y a de plus en plus de travaux. Généralement, sont développées des approches statistiques même si la tendance aujourd'hui consiste à les combiner à des analyses linguistiques souvent sommaires. Nous ne présenterons ici qu'une approche, le TextTiling de Hearst (1997), qui reste encore aujourd'hui la méthode la plus utilisée pour ce type de segmentation. Selon Hearst (1997), cette tâche de segmentation peut potentiellement s'intégrer dans des applications d'extraction d'information ou de résumé automatique. Elle décrit un algorithme à deux tâches principales : d'un côté, l'identification des *subtopic segments* et de l'autre le repérage des *subtopic shifts*. Elle travaille sur les paragraphes ou sur des ensembles comprenant plusieurs paragraphes. Au final, le *TextTiling Algorithm* a pour objectif de segmenter le texte en plusieurs blocs contigus, qui ne se chevauchent pas et qui sont cohérent thématiquement. Des scores sont ensuite attribués à ces blocs de texte, et c'est l'attribution de ces scores qui participe à la segmentation. Nous développerons dans la 3^{e} partie les diverses étapes de la segmentation telle qu'elle est envisagée dans le cadre du TextTiling mais également telle que nous l'utilisons.

3 Les SEDIS-ε dans le cadre d'un projet industriel

3.1 Présentation et Définitions

Nous définissons un SEDIS-ε comme un segment textuel susceptible de contenir une ou plusieurs informations présentant cette particularité de pouvoir évoluer dans le temps et/ou qui relativement à des besoins éditoriaux nécessiterai(en)t d'être réactualisée(s) (*cf.* Laignelet, 2006a, 2006b, 2006c). Une double distinction notionnelle nous permet de rendre compte partiellement de la complexité de la notion de « mise à jour » qui peut être envisagée tant du point de vue de la tâche réelle à laquelle elle fait référence que du point de vue du linguiste dont l'objectif est de décrire l'objet textuel auquel il réfère. Nous faisons donc une distinction sur deux plans. Le premier plan concerne la nature de l'information à mettre à jour, laquelle peut être à strictement parler une mise à jour ou bien une réactualisation. Dans le cas de la **mise à jour**, l'information n'est plus vraie ou ne s'est pas vérifiée (c'est souvent le cas lorsque l'auteur fait des prédictions sur un fait ou un événement). Dans l'exemple suivant, l'information *« Il n'existe pas à l'heure actuelle de vaccin contre le sida. »* ainsi que ce qui suit n'est potentiellement plus vrai au moment de lecture/réédition ou alors, étant donné un possible caractère prédictif, on est en droit de se demander si elle s'est ou non vérifiée.

La découverte du virus a permis la mise au point d'une méthode de dépistage [...]. On peut ainsi savoir qu'une personne est infectée longtemps avant que la maladie ne se déclare. Il n'existe pas à l'heure actuelle de vaccin contre le sida. Si les thérapies actuelles permettent d'améliorer sensiblement la durée et les conditions de vie du malade, aucune n'est capable d'éliminer le virus.

Figure 1 : Exemple d'une mise à jour

Dans le cas d'une **réactualisation,** les segments contiennent une information qui restera vraie dans l'absolu mais, en vue d'une ré-édition et d'une diffusion, les événements et dates associés doivent être modifiés pour faire référence à un moment plus proche du moment de lecture/réédition. Dans l'exemple qui suit, la valeur chiffrée *« 160 millions »* associée à *« en 2002 »* reste vraie, qu'on lise la fiche en 2003 ou en 2007. Cependant, il est fortement souhaitable de fournir de nouvelles informations et notamment de donner les chiffres pour l'année la plus proche de la date de réédition de la fiche.

L'organisation mondiale de la santé (OMS) estime, en effet, à 160 millions le nombre annuel de nouveaux cas dans le monde en 2002.

Figure 2 : Exemple d'une réactualisation

Le second plan fonde la distinction à un niveau plus textuel et oppose le SEDIS-ε minimal au segment d'interprétation. Les SEDIS-ε sont ainsi envisagés comme des segments textuels à granularité variable, ce qui nous permet de répondre au mieux aux exigences industrielles : en effet, il semble préférable pour le rédacteur chargé de la mise à jour d'avoir accès à la fois aux expressions locales à mettre à jour à proprement parler, et en même temps de bénéficier d'un contexte textuel suffisamment large et plus global pour être en mesure d'interpréter et de cibler rapidement ce qui nécessite une mise à jour. Dans la figure 3, nous pouvons voir un certain nombre de SEDIS-ε minimaux. Ils apparaissent dans cet exemple dans les petits cadres (ovales et rectangulaires). Leur taille varie du mot au syntagme et ils peuvent être de diverses natures. Dans certains cas, ils se confondent avec la notion d'indice : c'est le cas notamment des dates et des valeurs chiffrées. L'ensemble de l'extrait correspond à la notion de segment d'interprétation : c'est un segment discursif plus long que les précédents ; il s'agit ici de l'exemple en entier. La taille minimale de ce type de segment est la phrase mais ils peuvent aussi couvrir un ou plusieurs paragraphes, voire la partie entière.

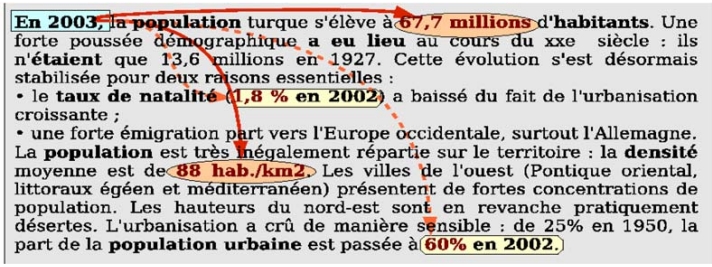

Figure 3 : Exemple présentant un cadre temporel ouvrant un segment d'interprétation

Dans cet exemple, le segment d'interprétation s'ouvre sur un introducteur de cadre temporel (Charolles, 1997). L'intérêt de considérer l'IC temporel *« En 2003 »* (dans le premier encadré) est que le critère sémantique (la référence temporelle *« 2003 »*) qu'il véhicule est valable pour

l'ensemble du segment donné. Ainsi, les deux valeurs chiffrées dans les ovales ont une relation (temporelle) à travers l'expression *« En 2003 »*. Les deux éléments dans les encadrés arrondis sont également des informations à mettre à jour du fait de leur proximité temporelle. Dans ce segment, il est important de noter que toutes les informations contenues ne sont pas à mettre à jour, par exemple *« Une forte poussée démographique a eu lieu au cours du XXe siècle [...] »,* pour lesquelles une référence temporelle différente est explicitement signalée. Nous définissons un segment d'interprétation comme un segment textuel de longueur indéterminée, présentant une certaine homogénéité sémantique (temporelle, spatiale, etc.), pouvant contenir des segments ne nécessitant pas de mise à jour et qui contient des SEDIS-ε minimaux et/ou des indices.

3.2 Repérage des SEDIS-ε : l'apport d'une méthode telle que le TextTiling pour le repérage des segments d'interprétation

Des travaux antérieurs menés dans une optique de segmentation ont cherché à combiner les marques de natures différentes, soit statistique et linguistique. Dans le cadre du projet REGAL, les auteurs cherchent à construire une structure thématique des textes afin de pouvoir soutenir une navigation intra-document dans le cadre de systèmes de résumé dynamique. Dans cette optique, Hernandez (2004) propose de combiner une analyse par segmentation lexicale avec un repérage de marques linguistiques. Cette étude nous semble très intéressante, à la fois de par la méthode utilisée mais également à travers les conclusions qu'elle apporte. Ainsi, selon Hernandez (2004 : 184), *« la cohésion lexicale apporte la robustesse au système en lui permettant de produire des résultats relevant de tout domaine [...]. Les marques linguistiques quant à elles apportent la finesse en permettant de repérer avec précision les bornes de segments, plus souvent la borne initiale d'ailleurs, la délimitation de la borne finale étant souvent très difficile voire impossible »*. Hernandez conclut en disant que *« la combinaison d'une analyse automatique par cohésion lexicale et d'un repérage des cadres apparaît comme triplement profitable : elle permet dans certains cas d'affiner l'ajustement des marques de segmentation automatique, dans d'autres de fournir un indice supplémentaire de fermeture de cadre ; enfin elle met en lumière un point important concernant la dimension lexicale des cadres de discours : il semble en effet que ceux-ci présentent une cohésion lexicale forte chaque fois qu'ils jouent un rôle procédural dans la classification des données transmises au lecteur. »* Dans le cadre de notre projet, nous cherchons à mettre en place une expérimentation telle que celle qui a été faite par Hernandez tout en étant conscient que notre objectif applicatif est bien différent. Ainsi, nous supposons que le repérage des SEDIS-ε peut être automatisé à travers la prise en compte d'indices linguistiques et discursifs, lesquels, bien qu'ils aient une fonction précise dans la langue, peuvent également permettre l'interprétation d'un segment comme étant de nature évolutive. S'ils sont considérés isolément, ces indices ne sont cependant pas suffisants pour répondre à cet objectif de repérage automatique des SEDIS-ε (Laignelet, 2006a) ; en revanche, envisagés en termes de configurations, *i.e.* en prenant en compte des combinaisons d'indices, il est tout à fait pertinent de penser que cette tâche peut être automatisée. En plus de la combinaison d'indices linguistiques et discursifs – et c'est l'objet de cet article – nous souhaitons analyser l'impact d'une analyse statistique de type TextTiling. Dans notre cas, et au stade de notre étude, nous souhaitons observer si les ruptures thématiques engendrées par une telle segmentation peuvent elles-mêmes devenir des indices pour le repérage des frontières initiales et/ou finales des SEDIS-ε de type « segments d'interprétation ».

4 Protocole expérimental

4.1 Annotation manuelle

Nous travaillons actuellement sur un corpus constitué de 92 textes de type encyclopédique. Il s'agit de fiches encyclopédiques éditées et accessibles sur le marché de l'édition (propriété des Editions Atlas). *A priori* du point de vue du type de texte (*cf.* terminologie de Biber), ce corpus est homogène. Le trait distinctif entre ces textes relève de la catégorisation en genre et plus précisément du domaine de connaissance auquel chacune des fiches appartient. Nous insistons sur ce point parce que nous supposons l'importance de cette distinction par domaine pour les résultats[3]. Ce corpus constitue une base de 80 000 mots, au format XML. Sur le total des 92 fiches, nous avons procédé à l'annotation manuelle de 38 d'entre elles. Cette tâche d'annotation manuelle a consisté à marquer à l'aide de balises XML les frontières des segments qui sont potentiellement des segments contenant de l'information à mettre à jour. Elle met en évidence la présence de 630 SEDIS-ε dans les 38 fiches parcourues. Par choix méthodologique[4], les SEDIS-ε sont de longueur égale à l'unité phrase (*i.e.* qui commence par une majuscule et se termine par une marque de ponctuation).

4.2 Outil : LinguaStream

Même si ce n'est pas l'objectif central visé par article, il est important de préciser que l'ensemble des marqueurs de surface[5] sur lesquels nous travaillons sont repérés de manière automatique à l'aide de la plateforme LinguaStream (*cf.* Widlöcher et Bilhaut, 2005). LinguaStream est une plate-forme générique pour le traitement automatique des langues qui permet d'effectuer des traitements et des analyses de types et de niveaux linguistiques variés (morphologique, syntaxique, sémantique, discursif ou encore statistique) sur des corpus en XML : il offre la possibilité d'utiliser différents langages en fonction de ce qu'on veut faire : des lexiques, des grammaires Prolog, des expressions régulières, des macro-expressions régulières, des programmes groovy, etc. . Travaillant sur des objets « mouvants » (qui peuvent aller de la taille d'une phrase à la taille d'une partie entière) cet outil est particulièrement pertinent de par les diverses possibilités de visualisation qu'il offre. Concernant cette étude, LinguaStream nous permet de travailler directement sur notre corpus annoté manuellement des SEDIS-ε, d'y adjoindre un découpage issu d'une segmentation TextTiling et de comparer, observer et analyser aisément les deux types de segmentation obtenus.

4.3 Détail des étapes de la segmentation par le TextTiling et adaptation à notre étude

L'algorithme du TextTiling permet une segmentation des textes fondée sur la notion de changement thématique. L'hypothèse de Hearst (in Hernandez, 2004 : 191, *note n°86*) est que *« un ensemble d'items lexicaux est utilisé pendant la discussion d'un sous-thème, et quand le*

---

[3] Huit domaines différents sont représentés : géographie (14), médecine & santé (13), sciences & techniques (10), société (8), sport (8), histoire (17), art & littérature (12) et faune & flore (7). Nous ne traitons pour cette étude que les quatre premiers domaines cités.

[4] Cela nous permet notamment de pouvoir nous baser sur des unités homogènes (la phrase) lors de l'évaluation de nos programmes.

[5] des adverbes de temps, des syntagmes nominaux de temps, des superlatifs, des noms propres, des sigles, des superlatifs (Laignelet, 2006a, 2006b, 2006c)

sous-thème change, une proportion significative du vocabulaire change aussi ». D'une manière générale, cet algorithme consiste à comparer des paires de passages successifs après pondération des mots de chacun des passages en fonction de critères de distribution et de co-occurrence lexicale. Après une première étape de tokenisation et d'étiquetage pour supprimer les mots vides susceptibles de parasiter les traitements ultérieurs, Hearst pose une taille fixe et arbitraire des passages de texte à comparer : elle définit les notions de pseudo-phrase (*token sequence*) et de pseudo-paragraphes (*block*). Ces deux éléments présentent la particularité d'être de taille homogène, les pseudo-paragraphes comptant un nombre donné de pseudo-phrases. L'auteur explique qu'elle a choisi de ne pas utiliser les phrases comme unités car la phrase est un segment dont la définition pose problème et de plus, le fait de comparer des segments de même taille rend cette comparaison plus aisée. Une fois les blocs définis, des scores de similarité vont être calculés entre chaque paire de blocs adjacents (cf. cosinus, coefficient de Dice, Jacquard, etc.) basés sur la fréquence des tokens de chaque bloc. Dans une dernière phase, les frontières des segments thématiques sont détectées par comparaison des différences scores qui ont été attribués à chacun des segments lors de l'étape précédente. Dans le cadre de notre expérimentation, nous avons fait le choix de travailler sur des segments dont la taille fait trois phrases « réelles »[6]. Nous verrons que ce choix n'est pas sans poser de problème dans des textes où la mise en forme matérielle est très riche. Nous avons exclu de considérer l'unité paragraphe du fait de la particularité de notre corpus : en effet, s'agissant de fiches encyclopédiques grand public et donc à fort impact visuel, les personnes chargées de la mise en page de ces fiches ne respectent pas ou peu la signification du saut de ligne. Nous souhaitions donc une méthode nous permettant à la fois de descendre en dessous du grain paragraphe, et de traiter des unités plus ou moins homogène en taille. Par ailleurs, nous ne prenons pas non plus en compte les titres dans les calculs de similarité. Nous avons utilisé le coefficient de Dice tout en faisant varier le seuil en fonction des thématiques des textes : pour les fiches *géographie*, le seuil est de 0,09 ; pour les fiches *médecine & santé*, il est également de 0,09 ; pour les fiches *sciences & techniques*, il est de 0,15 ; enfin pour les fiches *société*, il est de 0,17. Cette variation a été mise en évidence après divers tests effectués sur notre corpus ; ces différences reflètent ainsi les variations que l'on peut observer entre types et genre de textes.

5 Premiers résultats

Une fois que les différents traitements que nous venons de présenter ont été effectués sur notre corpus, nous avons observé et compté le nombre de fois où, d'un côté une balise ouvrante de SEDIS-ε correspond à une balise ouvrante de segment thématique TextTiling (« ouverture stricte »), et de l'autre une balise fermante de SEDIS-ε correspond à une balise fermante de segment thématique TextTiling (« fermeture stricte »). De plus, du fait que les blocs aléatoires utilisés pour faire les calculs TextTiling peuvent avoir des frontières un peu n'importe où, nous avons également observé et compté les cas où il y avait une phrase de décalage entre les deux types de borne. Le tableau suivant récapitule ces données :

[6] Par « réelles » nous entendons des phrases grammaticales qui commencent par une majuscule et se terminent par une ponctuation forte. Cela ne correspond donc pas à la notion de *pseudo-phrase* définie par Hearst.

	valeurs réelles		Pourcentage	
nombre de SEDIS-ε	630		100	
nombre de « fermetures strictes »	141	200	22,38 %	31,75 %
nombre de fermetures +/-1 phrase	59		9,36 %	
nombre de « ouvertures strictes »	114	172	18,09 %	27,30 %
nombre d'ouvertures +/-1 phrase	58		9,20 %	

Tableau 1 : Comparaison des frontières ouvrantes et fermantes des SEDIS-ε et des segments thématiques

Ces résultats nous encouragent fortement à considérer les ruptures thématiques issues d'une segmentation TextTiling comme des indices nous permettant d'automatiser le repérage de SEDIS-ε autant pour le repérage de leur borne initiale que pour celui de leur borne finale : en effet, 31,75 % des frontières finales de SEDIS-ε apparaissent simultanément avec une fin de segment thématique, et dans 27,30 % des cas, les frontières ouvrantes de SEDIS-ε et de segment thématique sont co-occurrents. La figure 4 montre un cas où l'on peut observer une co-occurrence à la fois des bornes initiales des deux types de segments et de leurs bornes finales.

<SEDIS-ε> <HEARST> En 2000, la Chine avait une production de pêche de capture estimée à 17 millions de tonnes. La France occupe le quatrième rang en Europe avec une production annuelle d'environ 600000 tonnes, poissons, crustacées et mollusques réunis. Les sources rapportées par la FAO (Food Agricultural Organization) font état d'une croissance du commerce halieutique international de 4 % par an, soit un montant évalué 55,2 milliards de dollar en l'an 2000. </HEARST> <SEDIS-ε>

Figure 4 : Exemple de co-occurrence entre SEDIS-ε et segment thématique

Ce qui est intéressant dans cet exemple 4, c'est qu'il montre également que d'autres indices peuvent être pris en compte comme les introducteurs de cadres temporels (ici *« En 2000 »*) (Charolles, 1997). Cela rejoint les conclusions de Hernandez (2004 :194) sur la relation entre cadre de discours et cohésion lexicale. Il apparaît donc indispensable de considérer la segmentation TextTiling non pas de manière isolée ou suffisante en elle-même, mais comme partie prenante de configurations d'indices de natures différentes (linguistiques et discursives). Mais les résultats que nous donnons dans le tableau 1 ne permettent pas de rendre compte d'un certain nombre de limites liées à la fois à notre méthode mais également liées à des difficultés inhérentes au type de corpus sur lequel nous travaillons. Nous venons de faire la remarque suivant laquelle il est nécessaire de considérer indices linguistiques et indice statistiques de manière complémentaire. En fait, la prise en considération de certains indices discursifs devrait être effective au moment de la délimitation des segments aléatoires : ceci permettrait par exemple dans l'extrait de la figure 5 de faire débuter le segment au moment où un introducteur de cadre est également présent. On peut alors supposer que la balise ouvrante <HEARST> se situerait simultanément avec l'ouverture d'un cadre temporel et dans ce cas précis avec celle d'un SEDIS-ε.

[...] Actuellement, la place très importante de la France au sein de l'ONU lui impose de répondre aux menaces régionales. <HEARST> Elle participe aux opérations de maintien de la paix sous l'égide de l'ONU. Des casques bleus français sont ou ont été présents en République Centrafricaine, en ex-Yougoslavie, à Jérusalem [...] </HEARST> [...]

Figure 5 : Limites de notre méthode : prendre en compte des indices discursifs comme les introducteurs de cadre de discours

Par ailleurs, nous avons déjà souligné le fait que les textes sur lesquels nous travaillons sont des textes dans lesquels la performance visuelle est prégnante. Ceci entraîne donc une mise en forme matérielle très riche que la segmentation aléatoire ne prend pas en considération. Concernant l'algorithme TextTiling, Hearst l'a évalué sur des textes descriptifs (*« expository texts »*) ce qui est très différent du type encyclopédique sur lequel nous avons travaillé.

 En revanche, certaines sectes [...], que les Renseignements généraux ont listées : - la déstabilisation mentale ; - le caractère exorbitant des exigences financières ; - la rupture induite avec l'environnement ;   - les atteintes à l'intégrité physique [...] ; - l'embrigadement des enfants ; - le discours plus ou moins anti-social ;   - les troubles à l'ordre public ; - l'importance des démêlés judiciaires[...]

Figure 6 : Limites de notre méthode : prendre en compte la mise en forme matérielle

Dans cet exemple, nous pouvons voir que la segmentation aléatoire se place n'importe comment au sein de l'énumération : nous proposons donc de modifier cette phase essentielle de segmentation du TextTiling en prenant en considération des indices relevant de la mise en forme matérielle. Il est ainsi pour les énumérations mais également pour les parties titrées courtes (de niveau 3).

6 Conclusion

L'hypothèse selon laquelle les frontières de segments thématiques peuvent servir d'indices contribuant au repérage des segments d'information évolutive semble se confirmer même s'il s'avère nécessaire d'approfondir les expérimentations présentées. L'objectif de ce travail n'est pas de montrer que les segments thématiques correspondent exactement aux SEDIS-ε mais il consiste à observer la possibilité d'exploiter les bornes initiales et/ou finales de tels segments conjointement à d'autres indices de natures diverses (adverbiaux de temps, superlatifs, temps verbaux, etc.) pour le repérage des SEDIS-ε. Ainsi, nous envisageons notamment d'adapter l'algorithme TextTiling et plus précisément la phase de découpage en pseudo-blocs avec des techniques de segmentation de nature linguistique telles que l'encadrement du discours ou encore les titres pour améliorer nos résultats et notre objectif général.

Références

CHAROLLES M. (1997). L'Encadrement du Discours, Univers, Champs, Domaine et Espaces. *Cahiers de Recherche linguistique.*

HALLIDAY M., HASAN R. (1976). *Cohesion in English.* Longman Group Limited, London.

HEARST M. (1994). Multi-paragraph segmentation of expository texts. Proceedings of *the 32th Annual Meeting of the Association for Computational Linguistics*.

HERNANDEZ N. (2004). *Description et Détection Automatique de Structures de Texte.* Thèse de doctorat, Université de Paris XI.

JACKIEWICZ A., MINEL J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. *Actes de la 10ème conférence sur le traitement automatique des langues naturelles, TALN.* Batz-sur-mer, France.

LAIGNELET M. (2006a). Repérage de segments d'information évolutive dans des documents de type encyclopédique. *Actes de la 13ème conférence jeune chercheur sur le traitement automatique des langues naturelles, RECITAL.* Presses Universitaires de Louvain, Louvain, Belgique.

LAIGNELET M. (2006b). Analyse discursive pour le repérage de segments d'information évolutive. *74ème Congrès de l'ACFAS, Description Linguistique pour le Traitemant Automatique du Français (DLTAF-ACFAS).* 16-18 mai 2006, Montréal, Canada.

LAIGNELET M. (2006c). Les titres et les introducteurs de cadre come indices pour le repérage de segments d'information évolutive. *Actes du Colloque International Discours et Document (ISDD'06)*, Presses Universitaires de Caen, France.

MARCU D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press.

MINEL J.-L. (2002). *Filtrage sémantique, du résumé automatique à la fouille de textes*, Hermès.

PERY-WOODLEY M.-P. (2005). *Discours, corpus, traitements automatiques.* Hermès.

ROSSI J., BERT-ERBOUL A. (1991). Sélection des informations importantes et compréhension de textes. *Psychologie Française*.

WIDLÖCHER A., BILHAUT F. (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus, Actes de la *12e Conférence Traitement Automatique du Langage Naturel (TALN).* Dourdan, France.