Repérage de structures thématiques dans des textes

Olivier Ferret, Brigitte Grau, Jean-Luc Minel, Sylvie Porhiel

Repérage de structures thématiques dans des textes

Olivier Ferret (1,4), Brigitte Grau (1), Jean-Luc Minel (2) et Sylvie Porhiel (1,3)

(1) LIMSI – CNRS
BP 13391403 Orsay Cedex
Brigitte.Grau@limsi.fr
(2) CAMS, équipe LaLIC – CNRS, EHESS, Université Paris-Sorbonne
96 Boulevard Raspail 75 006 Paris
minel@msh-paris.fr
(3) LATTICE – CNRS
ENS, 1 rue Maurice Arnoux – 92120 Montrouge
sylvieporhiel@hotmail.com
(4) CEA DTI/SITI
91191 Gif-sur-Yvette Cedex
olivier.ferret@cea.fr

Résumé – Abstract

Afin d'améliorer les performances des systèmes de résumé automatique ou de filtrage sémantique concernant la prise en charge de la cohérence thématique, nous proposons un modèle faisant collaborer une méthode d'analyse statistique qui identifie les ruptures thématiques avec un système d'analyse linguistique qui identifie les cadres de discours.

To improve the results of automatic summarization or semantic filtering systems concerning thematic coherence, we propose a model which combines a statistic analysis system identifying thematic breaks and a linguistic analysis system identifying discourse frames.

Mots Clefs : Cadre thématique, cohérence thématique, exploration contextuelle.

1 Introduction

La cohérence joue un rôle essentiel au niveau de l'argumentation et de la progression thématique. Toutefois, si la notion de cohérence a été abondamment commentée (Halliday et al., 1976, Charolles, 1995, 1997), elle a été peu prise en compte dans les systèmes de résumé automatique ou d'analyse thématique. En fait, lorsqu'il s'agit de repérer automatiquement les thématiques d'un texte et de prendre en charge la cohérence thématique, les systèmes de

résumé automatique se heurtent à toutes sortes de difficultés. Premièrement, les modèles qui intègrent et exploitent des connaissances ou des ressources linguistiques (Berri et al., 1996, Mitra et al., 1997) ne s'appuient pas sur une vision globale du texte et des thèmes abordés : ils se fondent sur la notion de saillance d'une unité textuelle, d'une phrase ou d'un paragraphe, et cette saillance est calculée indépendamment de la structure thématique du texte. Deuxièmement, ces systèmes ne répondent que partiellement aux besoins des utilisateurs : ce qui est pertinent pour les uns ne l'est pas pour les autres (Sparck Jones, 1993, Minel et al. 1997), notamment parce qu'un utilisateur peut être intéressé par une thématique qui n'est pas prise en charge directement par l'auteur.

Afin d'améliorer la performance des systèmes de résumé automatique ou de filtrage sémantique, nous proposons un modèle qui met en jeu deux idées essentielles[1]. D'une part, ce système, pour répondre aux attentes du maximum d'utilisateurs, doit reposer sur des indicateurs linguistiques indépendants des sujets abordés dans les textes traités, l'intégration de connaissances du domaine demeurant cependant possible. D'autre part, le système de fouille et de filtrage doit pouvoir fournir des extraits de texte en rapport avec la thématique intéressant l'utilisateur et donc tenir compte de la structure thématique du texte original.

Les améliorations décrites dans cet article portent sur deux points qui sont étroitement liés, à savoir : le repérage des unités thématiques (i.e. des segments fortement cohésifs) et la segmentation des données textuelles (i.e. le découpage des textes en unités de nature textuelle faisant référence à un seul thème). Nous présenterons tout d'abord deux approches qui repèrent des thématiques, puis nous décrirons comment nous modélisons les marqueurs linguistiques dans la plate-forme Filtext (développée au CAMS) au moyen du logiciel ContextO. Enfin, nous illustrerons sur un article du *Monde Diplomatique* les premiers résultats obtenus.

2 Deux approches pour repérer les thématiques

2.1 Une approche linguistique

Nous proposons de repérer les thématiques d'un texte dans la perspective textuelle de M. Charolles (1997). Nous utiliserons la notion de *cadres* pour désigner les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements. Parmi les deux grandes catégories qui instancient des cadres :

1. vérifonctionnels (univers de discours) qui précisent les circonstances dans lesquelles une ou plusieurs propositions peuvent être validées (cadre spatial : *En Corée du Sud* ; cadre temporel : *En 1990* ; cadre énonciatif : *Selon le président des États-Unis* ; cadre de représentation : *Dans le dernier film de Lelouch* ; cadre de connaissances *: Dans la perspective de la physique quantique*) ;

[1] Dans le cadre d'un projet de recherche auquel collabore le CEA, l'équipe LaLIC du CAMS, le LATTICE et le LIMSI. Ce projet est financé par l'Action Concertée Incitative Cognitique 2000.

2. thématiques qui précisent le thème de la ou des propositions qui suivent (*En ce qui concerne la Corée du Sud*).

Nous ne traiterons que les expressions linguistiques qui instancient des cadres thématiques comme : *au sujet de, à propos de, en ce qui concerne, au chapitre (de), concernant, sur, quant à*, etc. Parmi leurs propriétés, on retiendra que les introducteurs thématiques sont syntaxiquement peu soudés, qu'ils préfixent une ou plusieurs propositions et qu'ils partitionnent l'information. Ces expressions sont intéressantes dans la production d'un résumé car ce sont des marqueurs linguistiques indépendants du contenu sémantique des textes qui peuvent donc être réutilisés pour tout type de discours. En outre, l'auteur les emploie pour signaler aux lecteurs comment il organise les informations ; les expressions introductrices de cadres permettent aussi de mettre en valeur ce qui est important pour le rédacteur. Du point de vue du lecteur, ces introducteurs de cadres sont des balises, des guides qui soulignent une intention informationnelle de l'auteur du texte (Porhiel, 1998). Ils instaurent par conséquent un lien de cohérence, produit par le rédacteur et reconstruit par le lecteur, dont les introducteurs sont l'expression formelle.

Selon la nature sémantique des cadres qui apparaissent à la surface des textes, il est possible de reconstruire plusieurs stratégies textuelles, chacune d'entre elles reflétant, en partie, les intérêts du lecteur. Un texte peut ainsi être organisé en fonction de cadres temporels, spatiaux, thématiques. Chacun des cadres rassemblent des informations qui sont liées à un introducteur et qui donc présentent une certaine unité ou continuité. La mise en place de ces cadres répond à une stratégie du rédacteur qui répartit les informations dont il fait état dans des rubriques homogènes par rapport à un certain trait (Virtanen, 1992 parle de *strategic continuities*). L'étude de T. Virtanen sur les adverbiaux de lieu et de temps souligne que les continuités dans la stratégie textuelle « créent cohésion et cohérence et [qu'] en même temps, elles segmentent le texte en signalant des frontières textuelles de nature différente » (Virtanen 1992 : 129 ; notre traduction). Ces remarques peuvent s'appliquer à l'ensemble des cadres.

On comprend dès lors l'importance et la pertinence d'un découpage en cadres dans une analyse textuelle automatique. Ce découpage s'avère essentiel pour le filtrage car un segment extrait peut parfaitement être intégré dans un cadre introduit par une expression qui ne figure pas dans ce segment. Il y a donc tout intérêt à colliger et à classer les expressions potentiellement introductrices de cadres de discours.

2.2 Une approche fondée sur des critères statistiques

La deuxième approche se fonde sur deux méthodes d'analyse thématique de texte et vise à déterminer les endroits du texte où le sujet traité (le thème) change.

La première méthode d'analyse, fondée uniquement sur des critères statistiques ou numériques, part du constat que le développement d'un thème entraîne la reprise de termes spécifiques, notamment lorsqu'il s'agit de textes techniques ou scientifiques. La reconnaissance de parties de texte liées à un même sujet est alors fondée sur la distribution des mots et leur récurrence. Si un mot apparaît souvent dans l'ensemble du texte, il est peu significatif, alors que sa répétition dans une zone limitée est très significative pour caractériser le thème de cette partie de texte. Le principe général appliqué par les différents systèmes (Masson, 1995, Hearst, 1997) consiste à associer un vecteur de descripteurs à une zone de

texte, où les descripteurs sont typiquement les mots lemmatisés du texte et leurs valeurs, le nombre d'occurrences de ces termes dans la zone. Le produit scalaire de ces vecteurs permet ensuite de regrouper ou de séparer les zones qu'ils décrivent s'ils sont proches ou non.

La deuxième méthode d'analyse, quant à elle, utilise des sources de connaissances externes non dédiées. En généralisant la notion de répétition de termes, la cohésion thématique d'un texte se traduit par l'utilisation de termes faisant référence à une même entité par l'emploi de synonymes, d'hyponymes, de mots liés sémantiquement ou appartenant au même domaine. Afin d'introduire ces caractéristiques dans une analyse vectorielle d'un texte, les travaux décrits dans (Ferret et al., 1998) utilisent un réseau de collocations construit automatiquement sur un corpus d'articles de journaux afin d'enrichir les vecteurs décrivant un paragraphe. Le principe utilisé consiste à augmenter la valeur de descripteurs lorsqu'ils sont liés de manière significative dans le réseau à un descripteur du paragraphe. Cette modification des vecteurs tend ainsi à rapprocher des paragraphes comportant des mots liés.

Notre projet vise à exploiter la complémentarité des approches statistiques et de l'approche linguistique. Les deux méthodes statistiques offrent le moyen de segmenter thématiquement des textes de différents types en s'appuyant sur un critère de distribution lexicale. Afin d'affiner la décision de lier ou de séparer des segments, nous proposons d'utiliser la présence de marqueurs linguistiques conjointement à ce critère. Nous mettons ainsi en œuvre une analyse globale, qui considère la répartition du lexique sur l'ensemble du texte, et nous la complétons par une analyse locale, qui met en évidence des marques de segmentation et de structuration argumentative (points de liaisons et ruptures potentielles). En ce qui concerne le repérage des unités thématiques, il s'agira de faire collaborer des procédures de calcul prenant en compte des indicateurs lexicaux (les introducteurs thématiques), à même de fournir très rapidement des indications sur le thème d'un segment de texte (i.e. une unité textuelle renvoyant à un seul thème) et les changements thématiques d'un segment à un autre, avec des marqueurs linguistiques porteurs d'informations quant au rôle des segments du point de vue argumentatif ou discursif.

3 Modélisation des marqueurs linguistiques

3.1 La méthode d'exploration contextuelle

Pour détecter les introducteurs thématiques, nous utilisons la méthode d'exploration contextuelle (Desclés et al., 1997, Minel et al., 2001) qui identifie des connaissances linguistiques en les restituant dans leurs contextes et en les organisant en fonction de tâches spécialisées. Tout comme la théorie des cadres, cette méthode d'analyse s'inscrit dans une optique textuelle : elle utilise une notion de contexte textuel qui s'exprime dans une règle d'exploration contextuelle. Une règle d'exploration contextuelle définit un espace de recherche : il s'agit d'un segment textuel[2] toujours déterminé à partir de la présence d'un marqueur déclencheur, appelé indicateur (par ex. les introducteurs thématiques) ; des indices

[2] Un segment textuel est considéré comme une suite w_i ou l'$indice_i$ représente le rang de l'unité lexicale dans le segment. La position d'un marqueur est alors repérée comme un couple (k, l) représentant respectivement le rang de début et le rang final de ce marqueur dans le segment textuel.

complémentaires doivent en général être recherchés dans l'espace de recherche en vue de confirmer ou d'infirmer la valeur sémantique de l'indicateur repéré (par ex. position dans la phrase, déclencheur non suivi de certains noms, adjectifs, etc.). L'ensemble de ces connaissances linguistiques sont organisées dans un modèle conceptuel (Ben Hazez et Minel, 2000, Minel et al., 2001) et exploitées par la plate-forme logicielle ContextO.

3.2 L'analyse linguistique des introducteurs thématiques dans le cadre de la méthode d'exploration contextuelle

Nous avons organisé les connaissances linguistiques issues de l'analyse linguistique en tenant compte des contraintes de ce cadre d'analyse et des contraintes de la modélisation. Deux types d'information sont pertinents : les informations formelles et les indices complémentaires.

Les informations formelles relatives aux introducteurs thématiques ne sont pas applicables à l'ensemble des éléments de ce groupe, à l'exception de la casse. Chaque marqueur doit être considéré séparément : certains sont simples (*concernant*), d'autres sont composés *(en ce qui concerne)* ; certains acceptent des variations en nombre (*au chapitre (de), aux chapitres (de))* ou des variations aspectuelles (*pour ce qui est/était*), d'autres ont un paradigme de prépositions et de déterminants (*en/pour ce qui concerne)*, d'autres encore acceptent des insertions et sont donc des lexies discontinues (*en ce qui concerne particulièrement cette loi)*.

Les indices complémentaires, quant à eux, permettent de ne pas extraire des lexies ou, au contraire, permettent de renforcer leur reconnaissance dans les textes. Ce sont ou bien des indices généraux qui s'appliquent à tous les introducteurs thématiques ou bien des indices spécifiques qui ne s'appliquent qu'à un ou n indicateur(s) particulier(s). Les introducteurs thématiques sont syntaxiquement détachés et placés à l'initiale. Toutefois, ils peuvent aussi être précédés d'autres constituants (adverbes énumératifs, des cadres spatiaux, etc) (2) ou de signes typographiques (1), ce qui prouve que la caractéristique positionnelle est trop forte. Pour pouvoir être correctement repérées par le système, les lexies doivent être combinées avec des indices qui renforcent ou infirment leur reconnaissance, et ce pour éviter de repérer des lexies formellement identiques, dépendant d'un autre constituant (3) :

1) *Des scénarios de politique-fiction s'élaborent: à propos de Montréal, on évoque ouvertement Belfast, Sarajevo ou Jérusalem.*

2) *Enfin, en ce qui concerne ce catalogue de mesures pour garantir le développement durable, (...) les 178 délégations (...)*

3) *Je souhaite apporter des précisions en ce qui concerne l'article.*

Mais ces indices généraux ne suffisent pas toujours à reconnaître les lexies qui sont des introducteurs. Des indices spécifiques doivent s'ajouter aux indices généraux pour lever des ambiguïtés syntaxiques ou lexicales particulières à un petit nombre d'introducteurs. L'introducteur *au chapitre* illustre le cas des ambiguïtés lexicales. Pour être un introducteur thématique (*Au chapitre société, le président propose*), ce marqueur ne doit pas être suivi de certains adjectifs (*Au chapitre suivant, l'auteur affirme que*), de chiffres (*Et, au chapitre XX on peut lire*) ou de guillemets (*Au chapitre « Technologie », l'étude de l'OCDE affirme*).

3.3 Expression des règles d'exploration contextuelle

Les règles d'exploration contextuelle, indépendantes les unes des autres et organisées en tâches (Ben Hazez et Minel, 2000), sont d'abord exprimées dans un langage formel de type déclaratif puis traduites en Java. Pour la tâche de repérage des structures thématiques, nous avons construit sept règles qui repèrent les configurations textuelles suivantes : a) l'indicateur se trouve en position initiale ; b) l'indicateur se trouve dans un item d'énumération ; c) l'indicateur se trouve après un ou plusieurs groupe(s) de mots qu'il est possible de recenser exhaustivement ; d) l'indicateur introduit un énoncé thématique après un adverbe sélectif. Un premier test, effectué sur un échantillon de textes journalistiques (du *Monde Diplomatique)* de trois à six pages, nous a permis de dégager les résultats suivants :

- Le repérage des introducteurs (à partir de 70 occurrences de marqueurs) placés à l'initiale ou après un autre groupe de mots est satisfaisant. Les règles permettent également de ne pas repérer les phrases tronquées précédées d'un adverbe sélectif.

- Il est nécessaire de compléter les connaissances linguistiques acquises et les règles concernant le détachement des introducteurs doivent être affinées.

- Le système peut être confronté à des problèmes de typographie quand, dans les textes, les mots de début de paragraphe sont écrits en majuscules. La plate-forme ContextO est sensible à la casse et ce jeu de majuscules et de minuscules joue un rôle important. Le texte traité n'est pas systématiquement converti en minuscules car cette transformation inhiberait la reconnaissance des entités nommées.

- Enfin, dans sa version actuelle, ContextO n'intègre pas d'analyseur morpho-syntaxique, ce qui oblige parfois à accepter certaines indéterminations. Toutefois, de nombreux exemples montrent que l'expression de patrons morpho-syntaxiques comme par exemple « Introducteur + ADJ » sont beaucoup trop généraux.

4 Premiers résultats

L'extrait analysé est tiré d'un texte de quatre pages du *Monde Diplomatique* (voir annexe) et développe un des points qui ont favorisé la réussite économique de la Côte d'Ivoire et de la Corée du sud. Dans cet extrait, les cadres soulignent la cohérence et la cohésion du passage, et segmentent le texte en unités thématiques.

4.1 Analyse linguistique

Comme nous l'avons déjà souligné, les cadres sont des guides qui permettent une interaction entre cohérence et cohésion et qui participent à la dynamique du texte. Deux sortes de cadres, qui correspondent à deux stratégies textuelles, sont instanciés dans ce passage : des cadres thématiques *(en ce qui concerne (...), quant au (...), pour ce qui concerne (...)*) et des cadres spatiaux (*en Côte-d'Ivoire, en Corée du Sud*). Les premiers soulignent le passage d'une thématique à une autre, les seconds le passage d'un point particulier à un autre (ici pays). Les introducteurs thématiques assurent la cohérence et la cohésion générale du passage. Les trois thématiques qu'ils ré-introduisent ont été précédemment énumérées en termes identiques et

dans le même ordre. Les cadres spatiaux, eux, servent à illustrer localement les thématiques introduites par les introducteurs thématiques. Conformément à ce qui a été annoncé dès le début, il y en a toujours deux, ce qui assure la « cohérence illustrative » de l'ensemble du texte.

En ce qui concerne la segmentation, les expressions introductrices de cadre sont détachées en début de phrase. Les trois cadres thématiques apparaissent en début de paragraphe ce qui facilite la lecture car l'alinéa « signale au lecteur qu'il vient de traiter une unité de sens et, qu'il va passer à une unité ultérieure » (Bessonnat, 1988). Chaque introducteur thématique introduit ici une nouvelle unité thématique (et par là même ferme la précédente) qui correspond au paragraphe, ce principe n'étant cependant pas absolu car un cadre thématique peut en subordonner un ou plusieurs autres. Chaque point introduit par un introducteur thématique est ensuite développé au regard de deux exemples la Côte-d'Ivoire et la Corée du Sud introduits par des *en* qui instancient des cadres spatiaux. Ces derniers se délimitent l'un l'autre et sont subordonnés aux cadres thématiques du début de chaque paragraphe.

En conséquence, dans cet extrait, on a une double structuration et une double cohérence. Chacune correspond à une stratégie textuelle et est instanciée par des expressions linguistiques différentes exerçant un contrôle interprétatif spécifique : les expressions qui instancient des cadres sont des marques linguistiques de surface qui fournissent des indications sérieuses de marques de cohérence et de segmentation.

4.2 Analyse statistique

La figure 1 illustre les résultats de la méthode de segmentation décrite dans (Masson 1995)[3] sur le texte pris comme exemple. L'abscisse de chaque point identifie une frontière entre deux paragraphes et son ordonnée donne la valeur de cohésion entre les deux paragraphes. Plus cette valeur est élevée et plus les deux paragraphes sont susceptibles d'être liés sur le plan thématique. Nous avons fait figurer sur la courbe le seuil *SL*, valeur de cohésion au dessus de laquelle deux paragraphes sont jugés thématiquement liés, et le seuil *SC*, valeur de cohésion en dessous de laquelle deux paragraphes sont jugés thématiquement distincts. Ces seuils sont déterminés en fonction de la distribution des valeurs de cohésion (Cf. (Masson 1998) pour une présentation détaillée de leur calcul). Entre ces deux valeurs, la cohésion calculée est jugée plus faiblement significative et l'information résultant de la présence d'éventuels indices linguistiques est alors considérée comme prépondérante.

C'est le cas par exemple pour les paragraphes 12 à 14, qui illustrent l'intérêt des seuils fixés et la complémentarité avec l'analyse linguistique. Ces paragraphes abordent chacun un thème spécifique mais en évoquant les mêmes acteurs (*Côte d'Ivoire*, *Corée du Sud*, *firmes*), ce qui rejoint la « cohérence illustrative » du texte. L'analyse statistique trouve donc des liens entre ces paragraphes – en particulier entre le §13 et le §14 – mais en nombre restreint, d'où des valeurs de cohésion entre les seuils *SC* et *SL*. Les introducteurs de cadre situés au début de ces paragraphes (*en ce qui concerne*, *quant au*, *pour ce qui concerne*) permettent alors de dépasser

[3] Les méthodes décrites dans (Ferret et al., 1998) et dans (Hearst, 1997) donnent des résultats globalement similaires.

l'incertitude de l'analyse statistique et d'identifier les changements de thèmes.

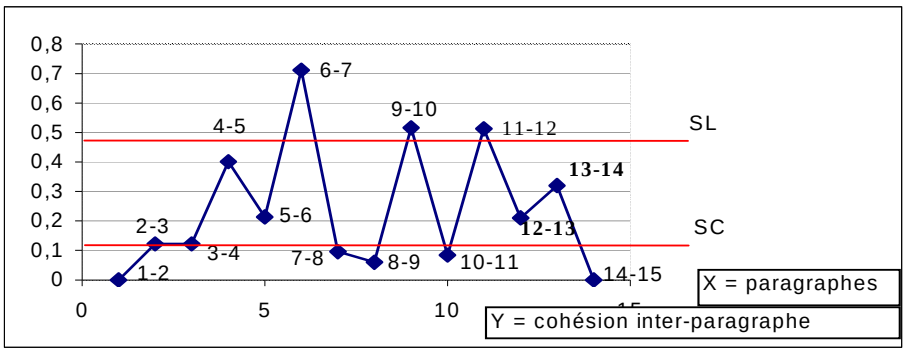

Figure 1 : Repérage de la structure thématique par l'analyse statistique

Plus globalement, la figure 1 montre que l'analyse statistique découpe le texte en grandes sections. Elle identifie les ruptures majeures, comme celle autour du paragraphe 8. Pour des coupures moins nettes, les informations fournies peuvent être confirmées par la présence de marqueurs. Nous travaillons actuellement à l'élaboration des algorithmes et des modèles qui nous permettront de généraliser l'intégration des deux approches.

5 Conclusion

Ces premiers résultats soulignent que la cohérence intervient à plusieurs niveaux et que les moyens linguistiques qui les expriment sont différents. Pour ce qui est de la prise en charge de la cohérence thématique, trois points nous semblent intéressants à développer ultérieurement. Premièrement, les conditions dans lesquelles les introducteurs thématiques renforcent les résultats de la segmentation thématique. Deuxièmement, les conditions dans lesquelles les introducteurs corrigent les résultats de la segmentation thématique. Des indices annoncent des thématiques et s'il était possible de repérer des phrases introductrices de thèmes (Cf. dernière phrase du § 11), la cohérence thématique des résumés en serait grandement améliorée. Enfin, il faudrait prendre en compte les différents niveaux de segmentation, en utilisant les expressions linguistiques instanciant des cadres à l'intérieur même des paragraphes.

Remerciements

Les auteurs remercient l'ensemble des participants qui collaborent à ce projet et qui ont alimenté par leurs différents travaux notre réflexion. 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.

Références

Ben Hazez, S., Minel J-L. (2000). Designing Tasks of Identification of Complex Patterns Used for Text Filtering. RIAO'2000, Paris, 1558-1567.

Berri, J, Cartier, E., Desclès, J-P., Jackiewicz, A., Minel, J-L. (1996). SAFIR, système automatique de filtrage de textes, *Actes du colloque TALN'96*, Marseille.

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

Charolles, M. (1997). L'encadrement du discours - Univers, champs, domaines et espace. *Cahier de recherche linguistique*, 6.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de linguistique*, 29, 125-151.

Desclés, J-P., Cartier E., Jackiewicz, A., Minel J-L. (1997*).* Textual Processing and Contextual Exploration Method In CONTEXT 97. Universidade Federal do Rio de Janeiro, Brésil : 189-197.

Ferret, O., Grau, B., Masson, N. (1998). Thematic segmentation of texts: two methods for two kinds of texts, *Actes ACL-COLING'98*, Montréal, Canada, volume 1 : 392-396.

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

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics, 23*, 1, 33-64.

Masson, N. (1998). *Méthodes pour une génération variable de résumé automatique : vers un système de réduction de texte*. Thèse de doctorat, Université Paris XI.

Masson, N. (1995). An automatic method for document structuring , *Actes 18th International Conference on Research and Development in Information Retrieval, ACM-SIGIR* , Seattle, USA, 372-373.

Mitra, M., Singhal, A., Buckely C. (1997). Automatic Summarization by Paragraph Extraction, *Proceedings of a Workshop : Intelligent Scalable Text Summarization*, EACL 97, Madrid, 39-46.

Minel , J-L., Nugier, S., Piat, G. (1997). How to appreciate the Quality of Automatic Text Summarization. *Workshop Intelligent Scalable Text Summarization*, EACL 97, Madrid, 25-30.

Minel J-L., Cartier, E , Crispino, G., Desclés J-P, Ben Hazez S, Jackiewicz, A. (2001) Résumé automatique par filtrage sémantique d'informations dans des textes. *Technique et Science Informatiques*, Paris, n° 3.

Porhiel, S. (1998). *Les introducteurs d'intérêt*. Thèse, Paris XIII. Diffusion, Presses du Septentrion.

Sparck Jones, K. (1993). What might be in a summary ?, in Knorz, Krause and Wormser-Hacker (eds.). *Information Retrieval 93*, 9-26, Universitates verlag Konstanz.

Virtanen, T. (1992). *Discourse Functions of Adverbial Placement in English*. Abo, Abo Akademi University Press.

Annexe

Les lexies en gras sont des introducteurs thématiques, celles en italique, des introducteurs spatiaux.

[§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.

[§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.

[§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.

[§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 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.

[§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.