Utiliser des corpus pour amorcer une analyse thématique

Olivier Ferret, Brigitte Grau

Utiliser des corpus pour amorcer une analyse thématique

Olivier Ferret — Brigitte Grau

LIMSI-CNRS
BP 133
91403 Orsay cedex
[ferret,grau]@limsi.fr

RÉSUMÉ. *L'analyse thématique est une étape importante pour de nombreuses applications en traitement automatique des langues, telles que le résumé automatique ou l'extraction d'information par exemple. Elle ne peut être réalisée avec une bonne précision qu'en exploitant une source de connaissances structurées sur les thèmes, laquelle est difficile à constituer avec une large couverture. Dans cet article, nous proposons de résoudre ce problème par un principe d'amorçage s'appuyant sur des corpus : une première analyse thématique, fondée sur l'utilisation d'une source de connaissances faiblement structurée mais aisée à construire à partir d'un vaste corpus – un réseau de collocations – permet d'apprendre, également à partir d'un corpus, des représentations explicites de thèmes appelées signatures thématiques. Ces dernières sont ensuite exploitées pour mettre en œuvre une seconde analyse thématique, plus précise et plus fiable.*

ABSTRACT. *Topic analysis is important for a lot of Natural Language Processing (NLP) applications, such as text summarization or information extraction. It can be achieved with a good precision only by using structured knowledge about topics, which is hard to obtain on a large scale. In this article, we tackle this problem by applying a bootstrapping mechanism that relies on corpora : a first topic analysis, which makes use of knowledge that is weakly structured but easy to build from a large corpus – a collocation network – permits to learn, also from a corpus, specific topic representations called topic signatures. These signatures are then used to support a second topic analysis that is more precise and more reliable.*

MOTS-CLÉS : *analyse thématique, segmentation thématique, amorçage, collocations, représentations thématiques.*

KEYWORDS: *topic analysis, topic segmentation, bootstrapping, collocations, topic representations.*

1. Introduction

La notion de thème est caractérisée par une forme de paradoxe. Elle est à la fois très intuitive et assez difficile à cerner. Tout le monde est capable d'en donner une définition informelle : le thème d'un texte, d'une conversation ou d'une partie de l'un ou de l'autre représente ce dont « parle » l'unité de discours considérée ; il en est le sujet. En revanche, les tentatives pour formaliser plus avant cette notion, aussi bien dans le champ de la linguistique que de l'analyse du discours, se sont avérées peu nombreuses.

Brown et Yule [BRO 1983] ont suggéré d'aborder ce problème en adoptant une méthodologie de type différentiel : ils proposent de caractériser les changements de thème plutôt que les thèmes eux-mêmes. L'analyse ainsi réalisée est appelée segmentation thématique puisqu'elle donne lieu à la délimitation d'une succession de segments de discours, individuellement homogènes sur le plan thématique. C'est la voie qui a été suivie par bon nombre de travaux en traitement automatique des langues, ainsi que nous le verrons au paragraphe 2.

Une autre approche de l'analyse thématique au niveau linguistique est incarnée par les travaux de Rastier [RAS 95], fondés sur la notion d'isotopie appliquée au niveau intra-textuel et inter-textuel. La notion d'isotopie renvoie dans ce cas à la récurrence au sein d'un texte ou d'un corpus d'une unité sémantique élémentaire, le sème. Selon cette perspective, un thème est défini comme une configuration stable de sèmes, récurrente dans un corpus. L'analyse thématique consiste alors à identifier de telles configurations dans les textes. C'est pourquoi nous parlerons à son propos d'identification thématique.

Les deux approches évoquées ci-dessus sont complémentaires : l'analyse des isotopies permet de caractériser le sujet abordé mais n'offre pas la possibilité de délimiter l'espace textuel concerné avec précision alors que la segmentation thématique possède les propriétés rigoureusement inverses. On retrouve à cet égard une distinction opérée classiquement en linguistique entre l'axe syntagmatique, occupé dans le cas présent par la segmentation thématique, et l'axe paradigmatique, incarné par l'identification thématique. Dans son extension la plus large, l'analyse thématique se doit donc d'intégrer ces deux axes au sein d'un même cadre.

Cette conception trouve son application aussi bien en linguistique que dans le traitement automatique des langues (TAL) et le travail que nous présentons dans cet article est une tentative pour la spécifier et la mettre en œuvre dans la perspective du TAL. Il repose sur une définition de la notion de thème adaptée à ce point de vue, assimilant un thème à l'ensemble des segments relatifs à ce thème dans un corpus. Cette définition porte en elle-même la démarche que nous proposons d'adopter pour développer une analyse thématique plus complète. Cette démarche se fonde dans un premier temps sur la segmentation thématique afin de construire une représentation des thèmes présents dans un corpus. Plus précisément, chaque représentation de thème rassemble et cumule l'ensemble des segments liés à ce thème dans le corpus considéré. Dans un second temps, ces représentations sont mises au service d'un processus intégrant à la fois segmentation et identification thématique.

Plus généralement, ce travail illustre l'intérêt de la notion d'amorçage dans le cadre d'un traitement automatique des langues s'appuyant sur des corpus : un processus d'analyse disposant de connaissances faiblement structurées est couplé à un mécanisme d'apprentissage automatique lui permettant de construire à partir d'un corpus des connaissances plus structurées ; celles-ci sont ensuite mises au service d'un processus d'analyse lui-même plus élaboré. Un amorçage de ce type permet ainsi d'améliorer et d'étendre les connaissances et les processus de façon incrémentale. Nous décrivons dans cet article la concrétisation de cette idée et son application à l'analyse thématique au travers du système ROSA.

Nous commencerons à la section 2 par un tour d'horizon des travaux sur l'analyse thématique, suivi à la section 3 d'une vue d'ensemble du système ROSA. Nous décrirons ensuite ses différentes composantes : la segmentation thématique reposant sur des connaissances faibles à la section 4, l'apprentissage des représentations de thème à la section 5 et l'analyse thématique les utilisant à la section 6. Nous exposerons ensuite les résultats de ce travail tout au long de la section 7, en passant du qualitatif au quantitatif pour finir sur une comparaison fine avec les résultats des travaux existants. La section 8 sera l'occasion de souligner quelques prolongements possibles.

2. Analyse thématique et TAL

Le parti pris de l'utilisation de l'amorçage dans le système ROSA est pour une part importante le résultat d'une volonté de réunir dans un même cadre les deux grandes approches qui ont été explorées dans le domaine de l'analyse thématique.

2.1. *Une approche fondée sur des connaissances structurées*

La première approche est incarnée par des travaux tels que [GRA 84] et [GRO 86], qui utilisent des connaissances de haut niveau à propos des domaines abordés afin de mener une analyse thématique très complète et très fine. Celle-ci permet de délimiter des segments thématiques, éventuellement non contigus, d'identifier le thème de ces segments mais également de déterminer la façon dont ils s'enchaînent et dont ils structurent un texte. Dans le cas de [GRO 86], cette analyse dépasse la simple dimension thématique pour s'inscrire plus largement dans l'analyse du discours. L'inconvénient évident de ce type d'approche réside dans l'étendue du travail de modélisation à accomplir, qui réserve pour le moment son application à des domaines restreints.

2.2. *Une approche quantitative*

La seconde approche est de nature quantitative et repose sur l'exploitation des informations de nature thématique que l'on peut extraire du niveau lexical pour

réaliser une analyse le plus souvent limitée à la segmentation des textes en unités adjacentes. Trois catégories peuvent être dégagées de ces études selon le type des ressources mobilisées.

2.2.1. *Exploitation des caractéristiques intrinsèques des textes*

La première catégorie regroupe des travaux s'appuyant seulement sur les caractéristiques intrinsèques des textes. La caractérisation des changements de thème est alors réalisée soit par un inventaire des marques linguistiques d'introduction d'un nouveau thème[1] ou de fin du thème courant[2] [PAS 97], soit par la détection d'une rupture de la cohésion lexicale, au sens défini par Halliday et Hasan [HAL 76]. Ce dernier cas est représenté par des systèmes tels que ceux décrits dans [HEA 97], [NOM 94], [MAS 95], [REY 94], [SAL 96] ou plus récemment [CHO 00], systèmes qui exploitent la façon dont les mots sont distribués dans les textes et prolongent ainsi sous des formes diverses le travail de Youmans [YOU 91], l'un des premiers menés dans cette optique. Compte tenu de la faiblesse des moyens déployés par cette première catégorie de méthodes, leur application est plutôt réservée à des textes possédant un vocabulaire spécifique et un style bien marqué, comme les textes techniques.

2.2.2. *Exploitation de connaissances sur la cohésion lexicale*

Un deuxième ensemble de méthodes fait appel à des connaissances externes aux textes que l'on peut considérer comme générales, c'est-à-dire non spécifiques des thèmes abordés. Nous classerons dans cette catégorie les travaux sur la segmentation thématique qui exploitent une source de connaissances cherchant à rendre compte de la notion de cohésion lexicale : un réseau de mots construit à partir d'un dictionnaire dans le cas de Kozima [KOZ 93], un thesaurus dans celui de Morris et Hirst [MOR 91] ou un réseau de collocations dans le cas de Ferret [FER 98a], [FER 98d] et de Kaufmann [KAU 99]. Ces méthodes ont montré leur efficacité sur les textes au sein desquels le vocabulaire est peu spécifique et où une même notion apparaît souvent sous de multiples formes, comme pour les textes narratifs. Cette efficacité est néanmoins conditionnée par la présence du vocabulaire employé dans le texte au sein du réseau lexical utilisé.

2.2.3. *Exploitation de connaissances thématiques*

Enfin, on trouve des méthodes s'inscrivant dans la problématique des modèles de langage probabilistes et qui sont à classer parmi les méthodes utilisant des connaissances thématiques spécifiques. Chaque thème pouvant être rencontré est

1. Ces marques sont généralement des mots ou des locutions apparaissant en tête de phrase comme maintenant, parce que, mais...
2. Il est à noter que les travaux de Passonneau et Litman relèvent partiellement de la première approche dans la mesure où ils font appel pour partie à une résolution des chaînes anaphoriques.

représenté par un modèle particulier, construit à partir d'un ensemble de textes sélectionnés manuellement pour leur représentativité de ce thème. Les réalisations menées dans le cadre de l'évaluation *Topic Detection and Tracking* (TDT) [FIS 99] relèvent majoritairement de cette approche[3]. Elle est également bien illustrée par le travail de Beeferman [BEE 99]. Dans les deux cas, la segmentation vise à séparer des textes – le niveau de granularité est donc assez élevé – et l'identification est relative à des thèmes très spécifiques, proches de la notion d'événement. Le système développé Bigi [BIG 98] montre néanmoins que des moyens globalement similaires sont utilisables pour segmenter le contenu des textes et identifier le thème des segments obtenus à un niveau beaucoup plus général.

2.3. *Notre positionnement*

Au regard de ces différents travaux, le système ROSA que nous présentons dans cet article constitue un système hybride. Il s'appuie en effet sur une méthode de segmentation quantitative utilisant une source de connaissances sur la cohésion lexicale afin de construire des connaissances thématiques spécifiques. Celles-ci lui permettent de mettre en œuvre une méthode d'analyse thématique alliant à la fois segmentation et identification thématiques. Le développement de cette démarche a pour objectif de se rapprocher progressivement des méthodes relevant de la première approche évoquée ci-dessus [GRA 84] [GRO 86] et d'aboutir à terme à une analyse précise des textes mettant en évidence leur structure thématique.

3. Vue d'ensemble du système ROSA

Le système ROSA (cf. figure 1) se décompose en deux sous-systèmes : SEGCOHLEX [FER 98a], qui segmente des textes en s'appuyant sur la notion de cohésion lexicale et SEGAPSITH [FER 98c] [FER 00], qui apprend incrémentalement des représentations de thème à partir des segments les plus cohérents de SEGCOHLEX et les utilise dans le cadre d'un module d'analyse thématique plus évolué.

SEGCOHLEX s'appuie sur une source de connaissances, en l'occurrence un réseau de collocations, construite automatiquement, non spécifique d'un domaine et faiblement structurée sur le plan thématique. Ce type de connaissances permet de développer un processus d'analyse réalisant une segmentation des textes avec des performances moyennes et sans identification des thèmes abordés dans les textes. Les segments de texte ainsi délimités, composés de mots pondérés, sont ensuite agrégés lorsqu'ils font référence à un même thème. Ce processus conduit à la production de signatures thématiques, elles-mêmes formées de mots pondérés. Ce nouveau type de connaissances, plus précises sur le plan thématique, permet la mise

3 Les systèmes dans le cas de TDT reposent souvent sur la combinaison de plusieurs types de modèles. Certains représentent effectivement des thèmes mais d'autres tentent de caractériser plus directement les changements de thèmes, notamment au travers de marques linguistiques.

en œuvre de l'analyse thématique de SEGAPSITH, capable d'identifier les thèmes des textes et de suivre une évolution thématique éventuellement non linéaire.

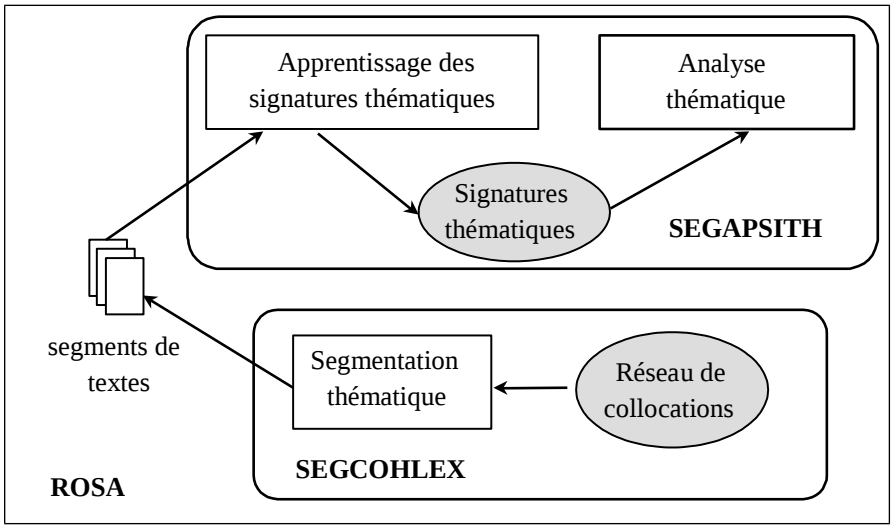

Figure 1. *Architecture du système ROSA*

Nous montrerons que les performances de ce second processus sont meilleures concernant la segmentation thématique que celles de SEGCOHLEX et plus globalement, qu'amorcer SEGAPSITH par SEGCOHLEX permet de construire des représentations de thème de meilleure qualité qu'en utilisant des textes non segmentés.

4. SEGCOHLEX

Nous débuterons la présentation de SEGCOHLEX par celle du réseau de collocations qu'il utilise en tant que référence concernant la cohésion lexicale.

4.1. *Le réseau de collocations de SEGCOHLEX*

Le réseau de collocations de SEGCOHLEX a été construit à partir d'un corpus composé de 24 mois du journal *Le Monde* répartis entre les années 1990 et 1994, ce qui représente un peu plus de 39 millions de mots. Compte tenu de notre cadre de travail, les textes ont été pré-traités de façon à ne retenir que leurs mots significatifs sur le plan thématique. Nous n'avons ainsi conservé que la forme canonique de leurs

mots dits pleins, c'est-à-dire les noms, dont les 2 300 noms composés[4] les plus fréquents sur 11 ans du *Monde*, les verbes et les adjectifs. Ce pré-traitement s'appuie sur le segmenteur *MtSeg* du projet *Multext* [VER 95] ainsi que sur l'étiqueteur morpho-syntaxique *TreeTagger* [STE 95]. L'extraction des collocations proprement dite a été réalisée par comptage dans une fenêtre glissante de 20 mots sélectionnés en adoptant la méthode décrite dans [CHU 90]. On trouvera dans [FER 98a] le détail de cette procédure d'extraction.

lemme1	lemme2	occurrences	cohésion
imprimante	ordinateur	13	0,227
bateau	voilier	125	0,224
prêtre	curé	44	0,209
policier	cambriolage	41	0,190
prendre	racine	120	0,110
collision	franc	7	0,076

Tableau 1. *Exemple des collocations extraites*

Le réseau obtenu est formé de 31 000 lemmes environ liés par quelques 7 millions de relations. Comme dans [CHU 90], nous avons adopté une estimation de l'information mutuelle comme mesure de la cohésion entre deux mots *x* et *y* :

$$coh(x,y) = \log_2\left(N \cdot \frac{f(x,y)}{f(x)\cdot f(y)}\right) \quad [1]$$

où *N* est la taille du corpus considéré ;
f(x,y) est le nombre de fois où *x* et *y* cooccurrent dans le corpus (dans l'espace de la fenêtre glissante) ;
f(x), respectivement *f(y)*, est le nombre d'occurrences de *x*, respectivement *y*, dans le corpus.

Dans [1], le rapport entre le nombre d'occurrences d'un mot ou d'un couple de mots et la taille du corpus permet d'estimer la probabilité du mot ou du couple de mots. Cette mesure de cohésion est ici normalisée par l'information mutuelle maximale relative au corpus, donnée par [2] :

$$I_{\max} = \log_2 N^2(Tf-1) \quad [2]$$

avec *Tf*, la taille de la fenêtre d'enregistrement et *N*, la taille du corpus.

Le tableau 1 donne quelques exemples de collocations couvrant le spectre des valeurs de cohésion et permettant de constater que ce réseau rend compte de relations entre les mots à la fois lexico-syntaxiques (prendre-racine), sémantiques

4. La sélection de ces noms composés a été réalisée grâce au logiciel INTEX [SIL 93].

(bateau-voilier) et pragmatiques (policier-cambriolage). Elle montre aussi l'existence de collocations beaucoup plus contingentes (collision-franc), présentes en grand nombre.

4.2. *La segmentation thématique de SEGCOHLEX*

La segmentation thématique de SEGCOHLEX est détaillée dans [FER 98a] et nous n'en donnerons ici qu'un descriptif assez bref. L'hypothèse qui sous-tend cette méthode est également celle qui fonde plus ou moins explicitement la plupart des travaux sur la segmentation thématique. Elle stipule que les ruptures de la cohésion lexicale dans les textes peuvent être assimilées à des changements thématiques. Nous lui adjoignons une hypothèse supplémentaire, affirmant qu'un réseau de collocations peut être utilisé pour détecter les ruptures de la cohésion lexicale.

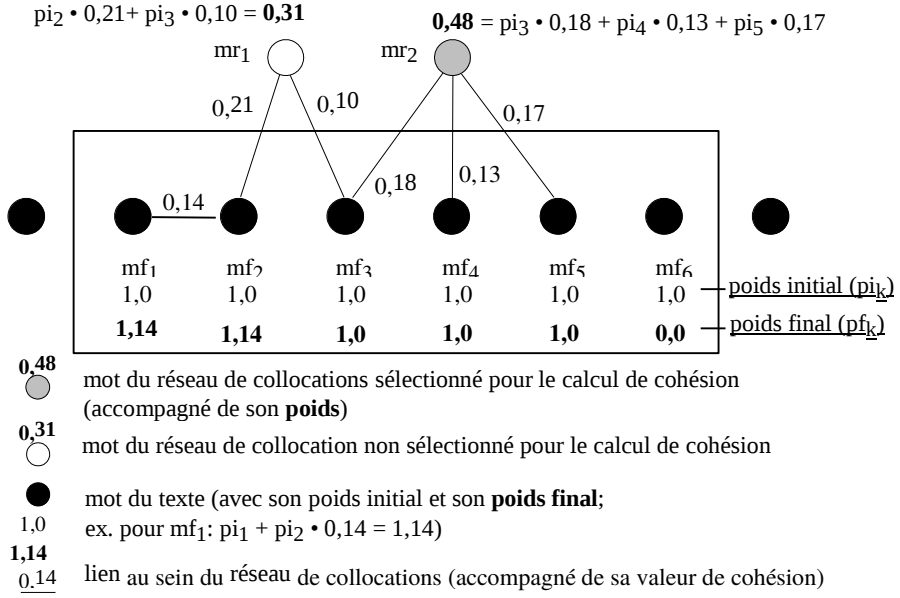

Figure 2. *Calcul du poids des mots*

À l'instar de [KOZ 93], l'évaluation de la cohésion d'un texte consiste ici à faire glisser une fenêtre d'une taille fixe sur ce texte et à calculer à chaque station de la fenêtre une mesure de la cohésion des mots qui y sont présents. Il faut préciser que les textes à segmenter ont auparavant subi le même prétraitement que celui employé lors la construction du réseau de collocations. La mesure de la cohésion des mots figurant dans la fenêtre s'appuie sur le réseau de collocations évoqué ci-dessus. Elle exploite l'hypothèse suivante : plus le nombre de mots de la fenêtre relatifs à un

même thème est grand et plus le nombre de liens que ceux-ci entretiennent dans le réseau de collocations, soit directement, soit par l'intermédiaire d'autres mots, est lui aussi grand. À l'inverse, lorsque les mots de la fenêtre relèvent de plusieurs thèmes, comme lorsque la fenêtre se trouve à cheval entre deux segments de thèmes différents, le nombre de liens trouvés dans le réseau entre ces mots est plus faible.

Ainsi que l'illustre la figure 2, cette densité en liens est caractérisée par le calcul d'un poids attribué aux mots de la fenêtre mais également aux mots du réseau de collocations liés à ceux de la fenêtre et sélectionnés de par leur proximité supposée avec eux. Cette proximité est déterminée par le nombre de liens qu'un mot du réseau entretient avec ceux de la fenêtre. Le poids d'un mot *m*, quelle que soit son origine, est égal à la somme des contributions de tous les mots m_i retenus auxquels il est lié, que m_i soit un mot de la fenêtre ou un mot du réseau sélectionné. La contribution de m_i est proportionnelle à la valeur de cohésion associée à la relation qu'il entretient avec le mot *m* dans le réseau de collocations.

Finalement, la mesure de la cohésion pour une position *p* du texte est donnée par la somme des poids ainsi calculés :

$$cohésion(p) = \sum_i signif(m_i) \cdot poids(m_i) \qquad [3]$$

où *poids*(m_i) est le poids du mot m_i (appartenant à la fenêtre ou ajouté à partir du réseau) calculé comme indiqué à la figure 2 ;
signif(m_i) est la significativité du mot m_i par rapport à un corpus de référence.

La significativité d'un mot se définit comme dans [KOZ 93] par l'information normalisée[5] de ce mot par rapport à un corpus, en l'occurrence le corpus de constitution du réseau de collocations. Elle est donnée par :

$$signif(m) = \frac{-\log_2(freq(m)/Tc)}{-\log_2(1/Tc)}, \text{avec } signif(m) \in [0,1] \qquad [4]$$

où *freq(m)* est le nombre d'occurrences du mot *m* dans le corpus de référence ;
Tc est la taille du corpus de référence.

La cohésion résulte donc de la combinaison, pour chaque mot impliqué, d'un poids dépendant de son contexte d'apparition et d'une mesure générale de sa capacité discriminante sur le plan thématique.

Son évaluation pour chaque position d'un texte permet d'obtenir la courbe globale de cohésion de ce texte. Celle-ci est d'abord lissée à l'aide d'une fenêtre permettant de moyenner localement les valeurs de cohésion. Les segments sont ensuite délimités par la détection des extrema de la courbe, un segment étant caractérisé par une séquence du type *minimum – maximum – minimum*.

5. Au sens de la théorie de l'information.

5. L'apprentissage de signatures thématiques dans SEGAPSITH

5.1. *Unités et signatures thématiques*

Le processus de segmentation de SEGCOHLEX permet de délimiter des zones de texte thématiquement homogènes servant ensuite de base à la construction d'Unités Thématiques (UTs). Une UT est la représentation d'un thème élaborée à partir d'un texte et incarne à ce titre un point de vue partiel sur ce thème. Une signature thématique rassemble ces différents points de vue pour un thème donné afin d'en construire une représentation plus générale et plus complète à l'échelle du corpus considéré. L'apprentissage de ces signatures thématiques [FER 98c] s'effectue par l'agrégation des UTs issues des textes de ce corpus au fur et à mesure de leur production, agrégation guidée seulement par une mesure de similarité. Il s'agit donc d'un processus d'apprentissage incrémental et non supervisé.

Concrètement, une UT est formée des mots du réseau de collocations les plus fréquemment sélectionnés pour le calcul de la cohésion lors de la délimitation d'un segment. Ce sont les mots du réseau les plus en rapport avec le segment et donc avec le thème auquel il fait référence. En pratique, ces mots sont retenus s'ils interviennent dans le calcul d'au moins 75 % des valeurs de cohésion se situant à l'intérieur du segment considéré. Par ailleurs, tous les segments ne conduisent pas à la formation d'une UT : seuls les segments dont la valeur de cohésion moyenne est suffisamment importante, c'est-à-dire les segments jugés véritablement cohérents sur le plan thématique, donnent lieu à la construction d'une UT. L'utilisation de mots sélectionnés à partir du réseau de collocations plutôt que des mots provenant des textes puise sa justification dans notre volonté de réduire le plus possible le bruit au sein des signatures, *i.e.* les mots sans rapport spécifique avec le thème représenté par la signature, ainsi que de s'affranchir de la forme d'expression des textes et de gagner ainsi en généralité. Ces mots sélectionnés sont pondérés au sein des UTs en tenant compte uniquement d'un indicateur général de spécificité, en l'occurrence leur significativité (cf. [4]).

Les signatures résultant de l'agrégation d'un ensemble d'UTs, leur structure est comparable à celle des UTs : chaque signature est un ensemble de mots pondérés. Seule la signification du poids des mots change d'un type d'entités à l'autre : il caractérise son niveau de généralité par rapport à un corpus de référence dans le cas des UTs alors qu'il traduit son importance vis-à-vis du thème représenté par la signature à laquelle il appartient dans le cas des signatures. Cette importance est évaluée pour le mot m_i par rapport à la signature $sign_j$ par la formule suivante :

$$poids(m_i, sign_j) = \frac{nbOcc(m_i, sign_j)}{nbAgr(sign_j)} \cdot signif(m_i) \cdot \frac{nbAgr(sign_j)^4}{(nbAgr(sign_j)+1)^4} \quad [5]$$

où *nbOcc(m_i, $sign_j$)* est le nombre d'occurrences du mot m_i dans la signature $sign_j$;

nbAgr(sign$_j$) est le nombre d'UTs rassemblées par la signature *sign$_j$*, égal au nombre d'agrégations ayant permis de la former.

Le premier facteur rend compte de l'importance du mot par rapport à la signature, le deuxième intègre la spécificité hors contexte du mot tandis que le dernier est un modulateur évitant que les signatures récemment créées se trouvent trop favorisées vis-à-vis des plus anciennes.

mots	occ.	poids	mots	occ.	poids
juge_d'instruction	58	0,501	chambre_d'accusation	47	0,412
garde_à_vue	50	0,442	recel	42	0,397
bien_social	46	0,428	présumer	45	0,382
inculpation	49	0,421	police_judiciaire	42	0,381
écrouer	45	0,417	escroquerie	42	0,381

Tableau 2. *La tête d'une signature sur le thème de la justice après 69 agrégations*

Le tableau 2 donne les mots les plus représentatifs, *i.e.* de plus fort poids, d'une signature sur le thème de la justice avec pour chaque mot son nombre d'occurrences, c'est-à-dire le nombre d'UTs dans lesquelles il apparaît, et son poids.

5.2. *Construction des signatures thématiques*

Les signatures thématiques sont formées par un mécanisme de regroupement incrémental des UTs intervenant à chaque mémorisation d'une nouvelle UT. Le processus se déroule selon un processus que nous évoquerons assez brièvement mais dont on peut trouver les détails dans [FER 98c].

Lorsqu'une nouvelle UT a été construite, on recherche en mémoire les signatures susceptibles de s'agréger avec elle en réalisant l'équivalent d'un pas de propagation d'activation. Nous sélectionnons ensuite les signatures les plus activées. L'activation d'une signature s'effectue par détermination des mots communs entre la signature et la nouvelle UT et tient compte du poids de ces mots à la fois par rapport à la signature (premier facteur) et par rapport à l'UT (second facteur) :

$$activ(sign_i) = \sum_j poids(m_j, sign_i) \cdot poids(m_j, UT) \qquad [6]$$

Seuls les mots ayant un poids supérieur à un seuil fixé (égal à 0,1) sont pris en compte pour évaluer l'activation des signatures, les autres étant majoritairement considérés comme du bruit. La sélection des signatures les plus activées s'effectue quant à elle par comparaison avec un seuil fonction de la distribution de ces

activations : seules les signatures ayant une activation supérieure à la somme de la valeur moyenne des activations et de leur écart type sont retenues.

Ce processus de sélection peut être vu comme une première mesure de similarité, peu élaborée mais applicable largement du fait de son coût raisonnablement faible. Après que le champ des possibilités a été restreint par cette première sélection, il devient possible d'appliquer une mesure de similarité plus complexe afin de déterminer si la nouvelle UT s'agrège à l'une des signatures sélectionnées. Cette mesure de similarité se fonde uniquement sur les mots communs entre une signature et une UT dans la mesure où la méthode d'apprentissage est source d'un bruit important, même si notre objectif est de le réduire le plus possible. En effet, les relations présentes dans le réseau de collocations ne sont pas seulement thématiques. Le type des collocations restant implicite, des mots sont retenus sur la base d'autres critères que la proximité thématique et représentent une source de bruit de notre point de vue. La mesure de similarité entre une UT et une signature combine l'importance que revêtent pour chacune de ces deux entités leurs mots communs par rapport à l'ensemble des mots qui les forment. Cette importance est évaluée à la fois en termes de poids et de nombre d'occurrences. On évite ainsi d'avoir une forte similarité entre une UT et une signature ne partageant qu'un petit nombre de mots communs de fort poids de part et d'autre. La combinaison de ces dimensions est réalisée par le biais d'une moyenne géométrique selon la formule suivante :

$$ratio_{\{s,u\}} = \sqrt{\frac{\sum_c poids(m_c, \{s,u\})}{\sum_t poids(m_t, \{s,u\})} \cdot \frac{\sum_c nbOcc(m_c, \{s,u\})}{\sum_t nbOcc(m_t, \{s,u\})}}$$
$$similarité(s,u) = \sqrt{ratio_s \cdot ratio_u} \qquad [7]$$

où l'indice *c* fait référence aux mots communs à l'UT (*u*) et à la signature (*s*) tandis que l'indice *t* désigne l'ensemble des mots constituant respectivement l'UT et la signature. Comme pour l'activation des signatures, les mots possédant un poids trop faible dans les signatures (poids inférieur à 0,1) ne sont pas retenus pour le calcul de similarité car ils sont globalement assimilés à du bruit.

Pour déterminer si une UT et une signature sont similaires, on compare la valeur de cette mesure de similarité à un seuil fixé *a priori*. Si la mesure est supérieure à ce seuil, l'UT et la signature sont supposées suffisamment proches sur le plan thématique et l'UT est agrégée à la signature. Dans le cas contraire, l'UT donne lieu à la création d'une nouvelle signature ne rassemblant dans un premier temps qu'elle-même. Ce seuil a été fixé à 0,25 dans le cadre des expérimentations que nous avons menées avec le souci de trouver un compromis acceptable entre une concentration des UTs dans quelques signatures et leur dispersion dans une multitude de petites signatures.

L'opération d'agrégation d'une UT et d'une signature est très simple puisque ces entités ont la même structure. Elle consiste pour l'essentiel en une fusion de deux

listes de mots pondérés : les mots de l'UT déjà présents dans la signature voient leur nombre d'occurrences augmenter d'une unité, et donc leur poids se renforcer, tandis que les autres sont ajoutés à la signature et viennent enrichir la description du thème.

Cette méthode d'apprentissage conduit à former des représentations de thèmes assez spécifiques, par opposition à des méthodes telles que celle présentée dans [LIN 97] par exemple, qui ont plutôt pour vocation à construire des représentations de thèmes très généraux, tels que l'économie, le sport, etc. Le travail de Lin se différencie également du nôtre dans la mesure où il dépend d'une classification *a priori* des textes. Celui de Pichon et Sébillot [PIC 00] est sur ce plan plus proche dans la mesure où il partage notre volonté de faire émerger des représentations de thèmes à partir d'un corpus de façon non supervisée. En revanche, le point de vue qu'adopte ce travail est différent puisqu'il se centre sur les mots plutôt que sur les segments pour construire sa classification. Plus précisément, il définit des représentations de thème en rassemblant des mots en fonction de leur distribution parmi les paragraphes des textes. Cette approche exploite l'information apportée par une vision d'ensemble du corpus considéré mais ceci, au détriment de l'incrémentalité. Ce point constitue un problème lorsque l'on souhaite construire des bases de représentations de thème facilement extensibles, comme c'est le cas ici.

5.3. *Une expérimentation de l'apprentissage de signatures thématiques*

Nous avons appliqué le module d'apprentissage de SEGAPSITH sur un mois (mai 1994) de dépêches de l'*AFP*. Les 7 823 UTs construites par SEGCOHLEX à partir de ces dépêches ont conduit à la formation de 1 024 signatures thématiques, dont 570 rassemblaient au moins 2 UTs. La signature du tableau 2 est l'une des signatures ainsi formées, rassemblant en l'occurrence 69 UTs relatives au domaine de la justice. Dans l'optique de leur utilisation par le module de segmentation de SEGAPSITH, nous avons sélectionné les signatures thématiques les plus fiables, à savoir celles possédant un nombre d'agrégations supérieur à 4. En mesurant la stabilité des signatures au fur et à mesure de leur formation[6], nous avons pu déterminer que le contenu d'une signature se stabilise en moyenne après une vingtaine d'agrégations. On observe néanmoins qu'après quatre agrégations, le niveau de stabilité atteint est déjà suffisamment important pour que les signatures obtenues puissent être utilisées. Cette valeur représente donc un bon compromis entre le nombre de signatures mises à disposition, donc l'étendue des thèmes pouvant être traités, et leur fiabilité.

Au sein des 193 signatures ainsi retenues, seuls les mots de poids supérieur à 0,1 ont été conservés afin de limiter le bruit, conformément au principe adopté lors du calcul de l'activation des signatures et de la mesure de similarité entre une UT et une

6. La stabilité d'une signature est évaluée par une mesure de similarité spécifique calculée à la suite de chaque agrégation la concernant entre son contenu avant et après l'agrégation (cf. [FER 98b]).

signature. Ce seuil a été fixé expérimentalement sur la base d'une étude manuelle des signatures construites.

6. L'analyse thématique de SEGAPSITH

Dans le prolongement des travaux portant sur la segmentation du discours tels que [GRO 86], l'analyse thématique de SEGAPSITH traite les textes linéairement et identifie les changements de thème sans différer sa décision, c'est-à-dire en tenant compte uniquement des éléments qu'elle a pu extraire de la partie du texte déjà analysée. Une fenêtre délimitant l'espace de focalisation de l'analyse est déplacée sur l'ensemble du texte considéré. Celui-ci a subi auparavant le même prétraitement que celui appliqué dans le cadre de SEGCOHLEX. Un changement de thème est détecté dès lors qu'une différence significative et durable est trouvée entre l'ensemble des signatures sélectionnées à partir de cette fenêtre de focalisation – ensemble définissant le contexte thématique de cette fenêtre – et l'ensemble des signatures associées au segment courant, ensemble formant le contexte thématique de ce segment.

6.1. *Notion de contexte thématique*

Le contexte thématique d'une entité a pour vocation de caractériser celle-ci sur le plan thématique. Concrètement, il prend la forme d'un vecteur de signatures thématiques. Chacune d'elles y est pondérée en fonction de son importance vis-à-vis des autres signatures du vecteur. La présence de plusieurs signatures s'explique en premier lieu par leur spécificité. Disposer d'un ensemble de signatures proches les unes des autres permet ainsi de couvrir un champ thématique plus large. Par ailleurs, SEGAPSITH manipule des représentations à base de mots et non de concepts. Le sens de ces mots restant ambigu, ceux-ci peuvent faire référence à plusieurs thèmes. Associer plusieurs signatures au sein d'un contexte permet de compenser et implicitement de lever les ambiguïtés existant à propos de certains mots.

6.1.1. *Contexte thématique de la fenêtre de focalisation*

Le contexte thématique de la fenêtre de focalisation est formé des *N* signatures thématiques les plus fortement activées par les mots de cette fenêtre, *N* étant la taille fixée pour tous les contextes. Le processus d'activation mis en jeu est équivalent à un pas de propagation d'activité. La valeur d'activation d'une signature thématique est donnée par :

$$activ(sign_i) = \sum_j poids(sign_i, m_j) \cdot nbOcc(m_j) \qquad [8]$$

où le premier facteur est le poids du mot m_j dans la signature $sign_i$ (cf. [FER 98c] pour plus de détails) et le second est le nombre d'occurrences de m_j dans la fenêtre de focalisation. Le poids d'une signature dans le contexte thématique est égal à sa valeur d'activation.

6.1.2. *Contexte thématique d'un segment*

Le contexte thématique d'un segment contient les signatures thématiques qui étaient les plus fortement activées lorsque la fenêtre de focalisation se trouvait dans l'espace de ce segment. Il est construit en fusionnant les contextes associés à la fenêtre de focalisation pour les différentes positions englobées par ce segment (cf. figure 3).

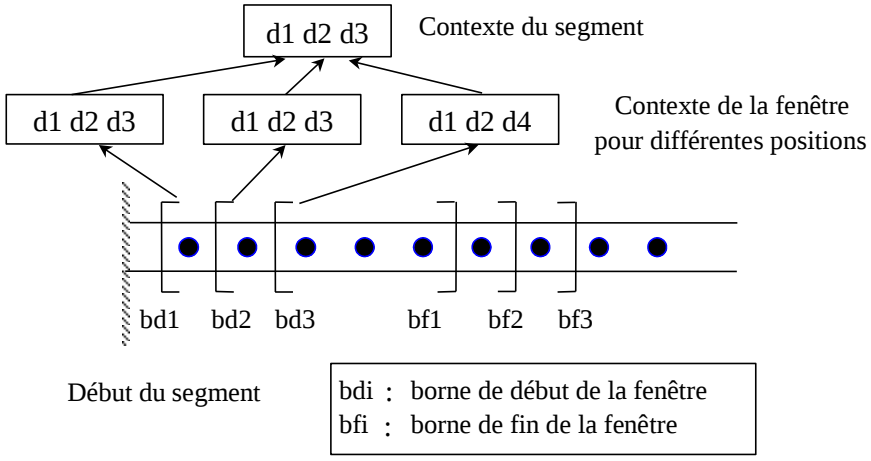

Figure 3. *Construction du contexte thématique d'un segment*

Cette fusion est réalisée incrémentalement : les signatures retenues pour chaque nouvelle position d'un segment sont réunies avec celles de son contexte courant. Les poids des signatures ainsi rassemblées sont ensuite réévalués selon la fonction suivante :

$$poids(sign_i, Cs, t+1) = \alpha(t) \cdot poids(sign_i, Cs, t) + \beta(t) \cdot poids(sign_i, Cf, t)$$

[9]

avec *Cf*, le contexte de la fenêtre de focalisation, *Cs*, le contexte du segment et $poids(sign_i,Cx,t)$, le poids de la signature $sign_i$ dans le contexte *Cx* pour la position *t*. Les résultats que nous présentons au paragraphe 6 ont été obtenus pour $\alpha(t) = 1$ et

$\beta(t) = 1$. Ces fonctions représentent un compromis concernant la vitesse d'évolution du contexte des segments. Cette évolution doit être suffisamment lente pour ne pas suivre de trop près les microvariations intervenant d'une position à une autre parmi les signatures activées, variations causées par la variabilité de l'expression de surface des textes. En effet, si le contexte d'un segment évolue à l'identique des contextes associés aux positions successives, il ne sera pas possible de détecter un éventuel changement de thème entre le segment courant et la nouvelle position de la fenêtre. À l'inverse, l'évolution du contexte des segments doit être suffisamment rapide pour ne pas introduire une distance trop importante entre la représentation stable du thème d'un segment et la façon dont ce thème est exprimé tout au long de celui-ci, distance qui risquerait d'entraîner la détection de faux changements de thème.

Après la réévaluation de leur poids, les signatures fusionnées sont classées par ordre décroissant de poids et seules les N premières sont conservées afin de former la nouvelle version du contexte du segment considéré.

6.2. *Similarité entre le contexte d'un segment et le contexte de la fenêtre de focalisation*

Afin de déterminer si le contenu de la fenêtre de focalisation est thématiquement cohérent avec le segment courant, le contexte thématique de la fenêtre est comparé au contexte thématique du segment. Cette comparaison s'appuie sur une mesure de similarité entre ces deux contextes, mesure tenant compte des quatre facteurs suivants :

— l'importance, en termes de poids, des signatures communes aux deux contextes ($sign_c$) par rapport aux signatures de la fenêtre (Cf) ;

— l'importance, en termes de poids, des signatures communes aux deux contextes ($sign_c$) par rapport aux signatures du segment (Cs) ;

— l'importance du nombre de signatures communes aux deux contextes par rapport à la taille d'un contexte. Ce facteur permet de s'assurer qu'une forte similarité ne sera pas trouvée entre deux contextes ne partageant qu'un petit nombre de signatures de fort poids (terme p/N dans [11]) ;

— la différence d'ordre parmi les signatures communes aux deux contextes.

Cette différence est donnée par :

$$diffRang(Cf, Cs) = \frac{\sum_{c=1}^{p} |rang(sign_c, Cf) - rang(sign_c, Cs)|}{(N-1)\cdot p} \quad [10]$$

avec p, le nombre de signatures communes, $sign_c$, une de ces signatures et $rang(sign_c,Cx)$, son rang dans le contexte Cx. La somme de ces différences de rang pour toutes les signatures communes est normalisée par une borne supérieure

correspondant à une situation hypothétique dans laquelle cette différence est maximale (égale à *N-1*) pour chacune des signatures communes.

Ces quatre facteurs sont combinés au moyen d'une moyenne géométrique (les quatre termes les représentant dans [11] apparaissent dans le même ordre) pour former la mesure suivante :

$$sim(Cf, Cs) = \left(\frac{\sum_{c=1}^{p} poids(sign_c, Cf)}{\sum_{i=1}^{N} poids(sign_i, Cf)} \cdot \frac{\sum_{c=1}^{p} poids(sign_c, Cs)}{\sum_{i=1}^{N} poids(sign_i, Cs)} \cdot \frac{p}{N} \cdot (1 - diffRang(Cf, Cs)) \right)^{1/4}$$

[11]

Le dernier terme est le complément du quatrième facteur, deux contextes étant d'autant plus similaires que leurs signatures communes se trouvent dans le même ordre. Finalement, deux contextes sont déclarés similaires si la valeur de [11] dépasse un seuil donné, fixé dans le cas présent à 0,5.

6.3. *Détection des changements de thème*

Sur le principe, l'algorithme de détection des changements de thème repose sur le calcul de la similarité entre le contexte thématique de la fenêtre de focalisation et celui du segment courant pour chaque position du texte considéré. Si cette valeur de similarité est inférieure à un seuil fixé initialement, l'algorithme en déduit la présence d'un changement de thème et un nouveau segment est ouvert. Sinon, le segment actif est étendu afin d'englober la position courante.

Cet algorithme de base fait l'hypothèse que la phase de transition entre deux segments est marquée de façon nette et sans ambiguïté. En réalité, la valeur de similarité entre contextes peut être localement fluctuante du fait de la forme de surface des textes. Il est donc préférable d'introduire un court délai avant de décider véritablement si le segment actif se termine ou si un nouveau segment s'ouvre. Pour tenir compte de cette incertitude, l'algorithme de segmentation prend la forme d'un automate (cf. figure 4) dont les transitions entre les quatre états sont contrôlées par les trois paramètres suivants :

— l'état courant de l'algorithme ;

— la valeur de similarité entre le contexte de la fenêtre de focalisation et le contexte du segment courant : *Sim* ou *non Sim* ;

— le nombre de positions successives de la fenêtre de focalisation caractérisées par un même état courant de l'algorithme : *nbConfirm*, qui doit être supérieur à $S_{confirm}$ pour sortir des états *DétectionNouveauThème* et *DétectionChangementThème*.

Au début d'un nouveau texte ou à la suite de la détection de la fin d'un segment, l'algorithme de segmentation se trouve dans l'état *ChangementThème*. Dès que le

contexte thématique de la fenêtre de focalisation demeure stable d'une position à une autre au regard de [11], il entre dans l'état *DétectionNouveauThème*. Il ne peut ensuite atteindre l'état *DéveloppementThème* que si cette stabilité est conservée pour les *nbConfirm - 1* positions suivantes. Autrement, il fait l'hypothèse qu'il s'agit d'une fausse alarme et revient à l'état *ChangementThème*. La détection de la fin d'un segment est le symétrique de la détection de son début. L'algorithme de segmentation entre en effet dans l'état *DétectionChangementThème* dès que le contexte de la fenêtre de focalisation change de façon significative entre deux positions successives. La transition vers l'état *ChangementThème* n'est cependant opérée que si ce changement se confirme pour les *nbConfirm - 1* positions suivantes.

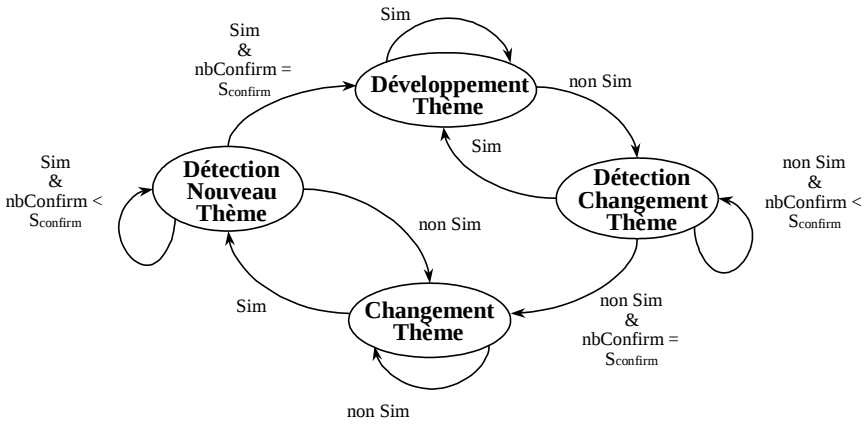

Figure 4. *Automate de détection des changements de thème*

Cet algorithme général est complété par deux mécanismes spécifiques. Le premier d'entre eux tient compte du fait que plusieurs segments d'un texte peuvent faire référence au même thème. Ce phénomène est intéressant à détecter car il constitue un premier pas vers la mise en évidence d'une structuration thématique complexe des textes. Pour ce faire, lorsque le segmenteur s'apprête à entrer dans l'état *DéveloppementThème*, il vérifie au préalable si le contexte du nouveau segment est similaire, selon [11], au contexte de l'un des segments déjà définis. Si une telle similarité est trouvée, le nouveau segment est lié au segment correspondant et adopte le contexte de celui-ci comme contexte propre. Le segmenteur fait ainsi l'hypothèse que le nouveau segment continue à développer le thème déjà abordé.

Le second mécanisme spécifique est lié quant à lui à l'état *ChangementThème*. Lorsque le segmenteur reste dans cet état pendant une durée trop longue (fixée à 10 positions de la fenêtre de focalisation dans nos expérimentations), il suppose que le thème évoqué par la partie de texte considérée n'est pas représenté parmi les signatures thématiques disponibles. Il définit alors un nouveau segment couvrant toutes les positions concernées et marque celui-ci comme ayant un thème inédit.

Bien entendu, ce mécanisme ne peut pas séparer plusieurs segments contigus faisant référence à des thèmes inédits. Néanmoins, il permet dans une certaine mesure de segmenter un texte sans disposer d'une représentation de tous les thèmes abordés par celui-ci.

7. Expérimentations et discussion

En préambule à l'exposé des expérimentations menées avec le système ROSA, il convient de souligner que les méthodes d'analyse thématique qu'il met en œuvre sont caractérisées, comme toutes les méthodes de nature quantitative, par un nombre important de paramètres. Des moyens d'optimisation existent pour fixer la valeur de ces paramètres afin de garantir de bonnes performances sur un corpus donné. Notre perspective n'est globalement pas celle-ci : ces moyens d'optimisation étant coûteux et nécessitant une référence, on ne peut en effet songer à les utiliser systématiquement. Nous préférons donc opter en faveur de valeurs de paramètres donnant des résultats ponctuellement moins performants mais plus uniformes sur différents corpus.

7.1. *Résultats qualitatifs*

Une première évaluation manuelle[7] de la méthode de segmentation de SEGAPSITH a été réalisée sur un jeu de test restreint sans faire intervenir de protocole formel tel que celui présenté dans [HEA 97]. Nous avons ainsi établi que pour le type de textes représenté par le texte de la figure 5, les résultats les plus intéressants sont obtenus avec une fenêtre de focalisation de 19 mots, une valeur de 3 positions pour le paramètre *nbConfirm* et des contextes d'une taille de 10 signatures. La figure 6 montre la valeur de similarité entre le contexte de la fenêtre de focalisation et le contexte du segment courant pour chaque position du texte donné. Les deux premiers changements de thème – le passage du concours de beauté au terrorisme et le retour au concours de beauté – sont clairement détectés comme le montre la forte baisse des valeurs de similarité autour des positions 62-63 pour le premier et 89-91 pour le second (reportées en gras dans le texte). En revanche, la méthode ne repère pas le dernier changement de thème (passage du concours de beauté à la manifestation féministe) du fait de la faible taille et de la faible spécificité de l'évocation du dernier thème.

Une jeune indienne remporte le titre de Miss Univers 1994

<ST> Un mannequin indien de 18 ans, Sushmita Sen, a créé la surprise samedi à Manille en remportant le titre de Miss Univers 1994, devançant deux

7. L'évaluation manuelle fait référence dans le cas présent à une confrontation des résultats obtenus avec notre propre jugement sur les textes considérés.

beautés sud-américaines, Miss Colombie, Carolina Gomez Correa, et surtout Miss Venezuela, Minorka Mercado, qui faisait figure de favorite du concours.
La jeune Indienne, une beauté brune aux yeux noisettes de 1,75 mètre, est la première candidate de son pays à remporter ce titre. Elle succède à Miss Porto Rico, Dayanara Torres, 22 ans, qui lui a remis sa couronne devant une audience télévisée estimée à 600 millions de personnes à travers le monde. Parmi les six finalistes, figuraient également Miss Etats-Unis, Frances Louis Parker, Miss Philippines, Charlene Gonzales, et Miss République Slovaque, Silvia Lakatosova.
Elles avaient été choisies parmi un groupe de dix demi-finalistes qui comprenait également les **représentantes** de l'Italie, de la Grèce, de la Suède et de la Suisse. </ST>
<ST> Quelques heures avant la cérémonie, un homme avait été tué par l'explosion d'un engin qu'il transportait à un kilomètre environ du Centre des Congrès où s'est tenu le concours de beauté, face à la baie de Manille. La police n'a pas pu établir immédiatement si cet incident avait un rapport avec le concours.
Jeudi, une bombe artisanale de faible puissance avait explosé dans une poubelle du Centre des **Congrès**, sans faire de dégâts. </ST>
<ST> La nouvelle Miss Univers, qui a remporté plus de 150.000 dollars de prix divers, a déclaré qu'elle se destinait au théâtre, à la publicité ou à l'écriture. Mais son vœu le plus cher, a-t-elle assuré, était de rencontrer Mère Teresa, parce qu'elle est "un exemple parfait d'une personne totalement dévouée, désintéressée et entière". </ST>
<ST> Alors que se déroulait l'élection, une centaine de féministes ont manifesté pacifiquement devant le Centre des Congrès, pour dénoncer le concours, affirmant qu'il servait à promouvoir le tourisme sexuel aux Philippines. </ST>

<ST>, </ST> : bornes des segments résultant des jugements humains

Figure 5. *Exemple de dépêche de l'AFP traitée*

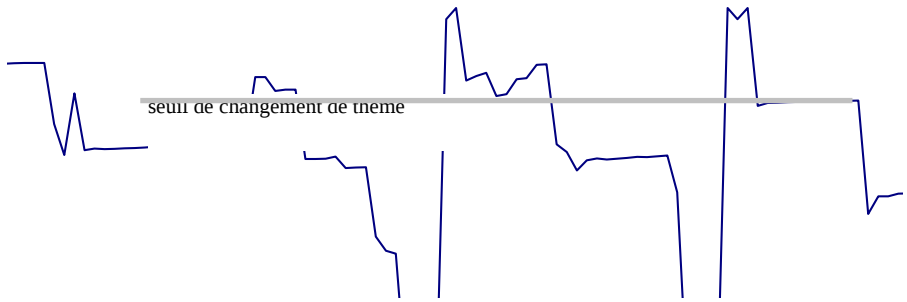

Figure 6. *Courbe de similarité entre contextes du texte de la figure 5*

Cet exemple illustre également deux caractéristiques importantes de notre méthode. Tout d'abord, l'utilisation d'une représentation explicite des thèmes permet de reconnaître que deux segments non contigus font référence au même thème. C'est le cas ici des segments 1 et 3 pour le thème du concours de beauté. Ensuite, notre méthode est capable de segmenter des textes sans avoir la représentation exacte des thèmes qu'ils abordent. Ainsi, la dépêche ci-dessus a-t-elle été segmentée en l'absence de signatures thématiques traitant spécifiquement des concours de beauté. Ce thème a été représenté par l'une de ses dimensions, en l'occurrence le fait qu'il s'agisse d'une compétition, au travers de la sélection de signatures concernant les compétitions sportives. Plus généralement, on voit donc que la représentation d'un thème évoqué dans un texte peut être construite dynamiquement par la composition de signatures faisant référence à une ou plusieurs de ses dimensions.

7.2. *Évaluation quantitative*

7.2.1. *Évaluation de ROSA*

De manière à évaluer l'intérêt du mécanisme d'amorçage, nous avons appliqué les méthodes de segmentation de SEGCOHLEX et SEGAPSITH sur une tâche classique d'évaluation des algorithmes de segmentation thématique : la redécouverte des bornes de textes dans un ensemble de textes concaténés. Nos méthodes de segmentation délimitant des segments du niveau du paragraphe, notre jeu de test est formé de textes courts, précisément 49 textes de 133 mots en moyenne provenant du journal *Le Monde*. Comme dans [HEA 97], la précision et le rappel sont définies par les rapports :

$$rappel = \frac{Nc}{D} \quad [12] \qquad précision = \frac{Nc}{Nb} \quad [13]$$

avec *Nb* : nombre de bornes trouvées par le système,
D : nombre total de limites de textes,
Nc : nombre de bornes trouvées correctes, *i.e.* qui correspondent à des limites de textes dans un intervalle de 9 mots autour de cette limite, après pré-traitement.

La f1-mesure, moyenne harmonique du rappel et de la précision, est utilisée de manière classique pour synthétiser ces deux mesures en un seul indicateur. Les résultats figurant dans le tableau 3 correspondent aux moyennes des valeurs obtenues sur 10 tests, l'ordre des textes ayant été modifié d'un test à l'autre.

Afin de disposer d'un point de référence minimal en termes de performances, nous avons implémenté une méthode de segmentation posant des bornes selon une procédure aléatoire. Par construction, une telle méthode impose de fixer *Nc*, le nombre des bornes posées. Nous avons choisi de prendre comme référence le nombre de bornes posées par SEGCOHLEX. Enfin, pour garantir le caractère véritablement aléatoire de la méthode, les résultats donnés sont des moyennes sur 1 000 applications[8]. Une première observation du tableau 3 permet de constater que tous les algorithmes en présence sont significativement meilleurs que cette méthode aléatoire (désignée sous le nom *Random* dans ce tableau).

méthodes de segmentation	rappel	précision	f1-mesure
Random	0.513	0,282	0.364
SEGCOHLEX	0,675	0,374	0,481
SEGAPSITH(1)	0,920	0,523	0,666
SEGAPSITH(2)	0,810	0,535	0,644

Tableau 3. *Résultat de l'évaluation des différentes méthodes de ROSA*

Concernant maintenant plus spécifiquement le système ROSA, les données de ce tableau montrent clairement que SEGAPSITH obtient de bien meilleurs résultats que SEGCOHLEX, ce qui tend à prouver l'intérêt effectif de l'utilisation de représentations explicites de thèmes par comparaison avec l'utilisation d'une source de connaissances plus « diluée » sur le plan thématique. La comparaison des deux versions de SEGAPSITH illustre également l'intérêt d'un amorçage par une première méthode de segmentation. SEGAPSITH(1) représente le système SEGAPSITH en tant que composante du système ROSA, avec un module

8. En pratique, on observe que les résultats se stabilisent assez rapidement à mesure que le nombre d'applications augmente et qu'il n'y a pas beaucoup de différence entre 100 applications et 1 000 applications.

d'apprentissage prenant comme entrée des UTs produites à partir des segments de SEGCOHLEX. SEGAPSITH(2) est une version de SEGAPSITH ne bénéficiant pas d'un amorçage par SEGCOHLEX. Plus précisément, chaque UT qu'il utilise pour former ses signatures est constituée de l'ensemble des mots pleins lemmatisés d'un texte. Ces UTs ne profitent donc ni du découpage en segments thématiquement homogènes ni de la sélection des mots d'un réseau de collocations en rapport avec les mots du segment. Or les données du tableau 3 montrent clairement que les performances de SEGAPSITH(1) sont meilleures que celles de SEGAPSITH(2). La légère baisse en précision est en effet largement contrebalancée par un rappel nettement supérieur.

7.2.2. *Comparaison quantitative avec d'autres méthodes*

Pour être complète, l'évaluation de ROSA doit être rapprochée des résultats d'autres systèmes similaires bien que cette comparaison ne soit pas toujours facile à mener du fait des différences existant entre ces différents travaux tant du point de vue des procédures d'évaluation que des jeux de test. Nous avons pris comme principal point de comparaison la méthode *TextTiling* [HEA 97]. Son statut de pionnière dans le domaine de la segmentation thématique et sa simplicité ont conduit d'autres travaux à en faire aussi un point de référence bien qu'il n'existe aucun corpus associé permettant une comparaison directe des différents résultats.

En ce qui nous concerne, nous avons réimplémenté la méthode *TextTiling* et nous l'avons évaluée sur le même jeu de test que ROSA, avec un pré-traitement identique des textes visant à ne sélectionner que les mots pleins des textes. Nous avons testé deux configurations de *TextTiling*, différant quant à la valeur des six paramètres caractérisant la méthode. La première configuration, TextTiling(1), correspond aux valeurs des paramètres données dans [HEA 97] : pseudo-phrases de 20 mots et blocs de 6 pseudo-phrases pour ne citer que les deux principaux paramètres. La seconde configuration, TextTiling(2), est le produit d'une optimisation des valeurs des paramètres vis-à-vis de notre jeu de test : pseudo-phrases de 10 mots et blocs de 10 pseudo-phrases. Les résultats sont donnés par le tableau 4. À titre indicatif, pour une évaluation similaire sur 44 textes (taille moyenne de 16 paragraphes), Hearst rapporte dans [HEA 97] un rappel de 0,95, une précision de 0,59 et une f1-mesure de 0,73.

méthodes de segmentation	**rappel**	**précision**	**f1-mesure**
TextTiling(1)	0.718	0.805	0.758
TextTiling(2)	0.806	0.847	0.826

Tableau 4. *Résultat de l'évaluation de TextTiling sur le jeu de test de ROSA*

On constate en premier lieu que les résultats obtenus par *TextTiling* sur notre jeu de test sont globalement meilleurs que ceux rapportés dans [HEA 97] alors que ces

derniers auraient plutôt laissé penser *a priori* que cette méthode était adaptée à des textes longs. Pour être précis, les performances en rappel sont plus faibles mais elles sont plus équilibrées par rapport aux performances en précision, ce que l'on peut considérer comme un meilleur cas de figure. Il est raisonnable de croire que dans le cas de [HEA 97], la granularité de la segmentation est inférieure à celle des textes, ce qui induit une tendance à poser davantage de bornes qu'il n'y a de limites de texte. Même avec un bon rappel quant à la détection de ces limites, la précision s'en trouve donc dégradée.

Il faut par ailleurs signaler que notre jeu de test est assez favorable à *TextTiling*[9] puisqu'il est constitué d'une succession de textes abordant des thèmes très différents. Dans une situation où les ruptures thématiques sont très nettes, comme c'est le cas ici de façon un peu artificielle, les méthodes s'appuyant sur la récurrence des mots telles que *TextTiling* se trouvent généralement avantagées par rapport aux méthodes utilisant une source de connaissances. Ces dernières sont en effet conçues pour trouver des liens plus profonds que les simples réitérations et tendent davantage à lier les parties de texte entre elles qu'à détecter leurs séparations.

Cette affirmation est corroborée par la comparaison des résultats de ROSA avec ceux de *TextTiling*. Dans le cas précis de SEGAPSITH(1), qui retient plus particulièrement notre intérêt, la présence de résultats assez semblables dans leur conformation avec ceux présentés dans [HEA 97] semble indiquer, conformément à l'analyse faite ci-dessus, que le niveau de granularité de la segmentation de SEGAPSITH est plus fin que celui de *TextTiling*, d'où un certain impact sur la précision de SEGAPSITH. Cette idée trouve d'ailleurs une confirmation dans les analyses qui sont faites du texte de la figure 5 par les deux méthodes. On se reportera au paragraphe 7.1 pour ce qui est de SEGAPSITH(1). Les deux configurations de *TextTiling* ne donnent quant à elle qu'une seule borne chacune. TextTiling(1) la place au début du deuxième segment (au niveau du mot « transporter ») tandis que TextTiling(2) s'avère un peu plus précise entre la plaçant à la frontière entre le premier segment et le deuxième segment (au niveau du mot « Suisse »). En fait, il apparaît que *TextTiling* détecte la présence d'un changement de thème important dans le texte et selon la valeur de ses paramètres, elle peut situer plus ou moins précisément où il se produit. En revanche, elle n'est pas capable de suivre la succession des changements thématiques[10] (en particulier le retour au premier thème mettant fin au deuxième segment) car les segments qu'ils délimitent ont une taille inférieure à la résolution que l'on peut intrinsèquement atteindre avec ce type de méthodes.

Pour achever cette comparaison avec *TextTiling*, il convient de souligner que les résultats donnés pour TextTiling(2) ont un caractère un peu artificiel puisqu'ils ont été obtenus au prix d'un processus d'optimisation qu'il ne serait pas réaliste d'appliquer pour tout texte à traiter. Ce processus a par ailleurs montré que la sensibilité de *TextTiling* à l'égard de la valeur de ses paramètres est assez forte et

9. Il est sans doute plus favorable que le jeu de test utilisé dans [HEA 97].
10. Il faut d'ailleurs rappeler que même SEGAPSITH(1) passe à côté du dernier changement de thème car on atteint là aussi la limite de sa résolution.

qu'il suffit souvent d'un écart faible pour obtenir des performances beaucoup moins bonnes. De ce point de vue, la stabilité des résultats de SEGAPSITH est plus grande, ce qui n'est pas complètement surprenant : lorsqu'un processus s'appuie sur des connaissances, il est parfois pénalisé par l'inertie qu'elles engendrent mais à l'inverse, elles le rendent beaucoup moins sensible aux caractéristiques des textes qu'il traite.

Compte tenu des différents points que nous avons relevés, il apparaît qu'une méthode de segmentation telle que celle de SEGAPSITH et une méthode de type *TextTiling* sont complémentaires. Comme nous l'avons déjà souligné dans [FER 98d], le problème n'est donc pas tant de choisir une fois pour toute une méthode « universelle » mais plutôt de savoir choisir la méthode la plus adaptée au contexte courant. Dans cette perspective, la segmentation de SEGAPSITH se présente comme un outil permettant d'opérer à un niveau fin de granularité et dans un contexte où la récurrence du vocabulaire n'est pas très importante. En outre, il convient de rappeler que cette segmentation est couplée à une identification des thèmes des segments, ce qui est une dimension de l'analyse thématique non couverte par les méthodes de type *TextTiling*.

Parmi les travaux réalisant l'intégration de ces deux fonctions, celui de Bigi [BIG 98] est sans doute celui qui se rapproche le plus du nôtre, bien que sa méthode d'évaluation soit légèrement différente. Dans ce cas, nous devrons nous contenter d'une comparaison avec les données fournies dans [BIG 98] puisque nous n'avons pas pu l'expérimenter sur notre jeu de test. La différence entre ses résultats – 0,75 en précision, 0,80 en rappel et 0,77 pour la f1-mesure – et les nôtres s'explique selon nous par la nature des thèmes abordés : son travail est fondé sur un petit ensemble de thèmes très généraux (affaires, politique, sport…) alors que nous considérons un large ensemble de thèmes spécifiques. Nos signatures thématiques étant plus proches des thèmes des textes, nous obtenons assez naturellement un meilleur rappel car le suivi des changements de thème est plus serré. Cette proximité explique également notre précision moindre : un suivi thématique plus fin conduit à distinguer davantage de segments ; or c'est une source de bruit du point de vue de la reconnaissance des bornes de texte lorsqu'on assimile texte et segment, ainsi que nous le faisons dans l'évaluation présentée au paragraphe 7.2.1.

7.3. *Quelques points de comparaison avec des travaux proches*

Une analyse thématique capable d'identifier le ou les thèmes d'un texte s'appuie toujours sur une représentation des thèmes, à l'instar de [BIG 98] ou des travaux effectués dans le cadre de l'évaluation *Topic Detection and Tracking* (TDT) [FIS 99]. SEGAPSITH exploite aussi des représentations explicites de thèmes mais en s'inspirant des systèmes fondés sur la cohésion lexicale, tels que ceux décrits dans [KOZ 93] ou [KAU 99], et non selon l'approche probabiliste généralement appliquée dans le cadre de TDT. Les travaux de TDT se différencient aussi des nôtres par le délai accordé pour décider d'un changement de thème : de 100 à 10 000 mots dans TDT pour seulement 3 mots pleins dans SEGAPSITH, ce qui a un impact

direct sur la taille moyenne des segments délimités. La tâche de segmentation de TDT est d'ailleurs clairement envisagée comme une tâche permettant de séparer des documents[11] et non comme une tâche ayant pour charge de les découper plus finement en fonction de leurs sous-thèmes.

Cette orientation conduit souvent à combiner plusieurs types de modèles de langage : certains sont véritablement de nature thématique mais d'autres sont plus spécialisés dans la caractérisation des frontières entre les documents. Or ces modèles atteignent souvent un degré de spécialisation vis-à-vis du corpus considéré que nous estimons être beaucoup trop important. Dans [BEE 99], Beeferman indique ainsi que les marques linguistiques trouvées comme les plus performantes pour retrouver les frontières d'un ensemble d'articles du *Wall Street Journal* sont tout à fait spécifiques de ce journal, comme le fait par exemple de trouver au début des articles le mot « incorporated » : les articles du *Wall Street Journal* contiennent en effet beaucoup de noms de sociétés et leur intitulé complet, qui comprend fréquemment le mot « incorporated », ne figure qu'au début des articles.

Dans SEGAPSITH au contraire, tout en nous appuyant sur des corpus, nous essayons d'atteindre un certain niveau de généralité. C'est ce qui nous a conduit par exemple à construire les signatures thématiques avec des mots sélectionnés à partir d'un réseau de collocations et non directement à partir des mots des textes. La dépendance vis-à-vis des spécificités propres aux textes traités est ainsi moins forte. Le même souci de généralité nous a guidés quant au choix de la nature des thèmes représentés. Alors que les thèmes de TDT représentent des événements et sont donc très spécifiques, que ceux de Bigi sont au contraire très généraux, nos signatures thématiques se situent entre ces deux extrêmes : elles visent à décrire des thèmes spécifiques, mais pas d'événements particuliers.

Les représentations de thèmes permettent à l'évidence de travailler à un niveau de granularité plus fin que la cohésion lexicale. Cependant, pour traiter des textes couvrant un large spectre thématique, il est nécessaire de construire ces représentations automatiquement, de préférence de manière non supervisée pour ne pas fixer de limitation relative aux thèmes attendus. Ce problème est partiellement abordé dans la tâche de détection de TDT, bien que ce ne soit pas son objet principal, mais il n'est pas lié à la tâche de segmentation. SEGAPSITH comprend au contraire un apprentissage non supervisé des représentations de thèmes qui sous-tendent son module d'analyse thématique. Son association avec SEGCOHLEX conduit en outre à amorcer ce module de segmentation à grain fin par un module plus fruste fondé sur la cohésion lexicale.

8. Conclusion et perspectives

Le développement de processus de traitement de la langue requiert de grandes bases de connaissances, très difficiles à produire. Afin de surmonter ce problème

11. Le terme « document » est à prendre au sens large puisqu'une partie des données est constituée de transcriptions d'émissions de radio ou de télévision.

pour l'analyse thématique, nous avons adopté une approche incrémentale fondée sur l'amorçage et l'utilisation de corpus : nous avons implémenté un premier processus d'analyse, fondé sur l'utilisation de connaissances acquises automatiquement à partir de corpus, utilisé ses résultats pour construire des connaissances plus structurées et enfin développé une seconde méthode d'analyse thématique exploitant ces connaissances afin d'obtenir de meilleurs résultats. L'évaluation de ce processus démontre que les résultats obtenus avec des connaissances spécialisées et structurées sont meilleurs que ceux obtenus avec des connaissances plus générales. Il montre, de plus, qu'apprendre des connaissances à partir des résultats d'une première analyse thématique fournit des connaissances plus fiables qu'apprendre à partir de textes non analysés.

Ce travail a permis d'illustrer de façon significative l'approche que nous proposions mais n'en couvre pas tous les aspects. Nous envisageons trois types d'extensions. Le premier concerne l'évaluation. Sur le plan de la segmentation, il est nécessaire de dépasser la tâche quelque peu artificielle de séparation d'un ensemble de documents pour s'orienter vers la segmentation des documents eux-mêmes. Ce type d'évaluation a déjà été mené (cf. [HEA 97] ou [PAS 97]) et repose sur la constitution d'une segmentation de référence par le recoupement des jugements émis par un ensemble suffisamment important d'annotateurs humains. Le cadre d'évaluation existe donc mais s'avère assez lourd à mettre en place. Contrairement au cas de la segmentation, l'évaluation de l'identification thématique de SEGAPSITH reste à faire. On pourra s'inspirer pour cela des procédures adoptées dans TDT concernant les tâches Detection et Tracking[12]. En revanche, un cadre d'évaluation réunissant segmentation et identification thématiques reste encore à définir.

Le deuxième axe des extensions à réaliser vise un changement d'échelle de l'analyse thématique de SEGAPSITH pour en faire une analyse largement utilisable, en particulier dans des applications de recherche et d'extraction d'information ne faisant pas de restriction sur les domaines abordés. On peut citer à cet égard l'exemple des systèmes de réponse à des questions factuelles tels que ceux conçus pour l'évaluation *Question Answering* de TREC. Pour que l'analyse thématique de SEGAPSITH puisse opérer dans un tel cadre, il est nécessaire de constituer une vaste base de signatures thématiques. L'objectif n'est pas de couvrir tous les thèmes imaginables mais plutôt de disposer d'une couverture thématique suffisante pour décrire chaque nouveau thème rencontré comme un assemblage de thèmes déjà présents dans la base des signatures. On s'appuie pour cela sur l'idée d'une compositionnalité possible des thèmes illustrée au paragraphe 7.1.

La dernière piste à explorer est liée au point que nous venons d'aborder. Pour être utilisable, une vaste base de signatures thématiques doit être dotée d'une structure plus complexe qu'une simple structure d'ensemble. Il faut donc étendre le

12. Le problème de l'identification thématique a été beaucoup moins traité en tant que tel que celui de la segmentation thématique, ce qui explique l'absence d'un cadre d'évaluation bien établi. À défaut, des travaux tels que [LIN 97] par exemple ont repris les méthodes d'évaluation utilisées en filtrage d'information, problème proche de l'identification thématique.

mécanisme d'apprentissage des signatures de SEGAPSITH afin qu'il produise une hiérarchie de signatures. Au-delà d'une dimension de structuration, cette hiérarchisation ouvre également la voie à des extensions fonctionnelles de l'analyse thématique de SEGAPSITH. Disposer de plusieurs niveaux de signatures doit en effet permettre de distinguer différents niveaux de thème au sein des textes et donc de rendre compte de leur structure thématique.

9. Bibliographie

[BEE 99] BEEFERMAN D., BERGER A., LAFFERTY J., « Statistical Models for Text Segmentation », *Machine Learning*, vol. 34, n°1-3, 1999, p. 177-210.

[BIG 98] BIGI B., DE MORI R., EL-BÈZE M. ET SPRIET T., « Detecting topic shifts using a cache memory », *Acte de 5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australie, 1998, vol. 6, p. 2331-2334.

[BRO 83] BROWN G.,YULE G., *Discourse Analysis*, Textbooks in Linguistics Series, Cambridge University Press, 1983.

[CHU 90] CHURCH K.W., HANKS P., « Word Association Norms, Mutual Information, And Lexicography », *Computational Linguistics*, vol. 16, n°1, 1990, p. 22-29.

[CHO 00] CHOI F., « Advances in domain independent linear text segmentation », *Actes de NAACL'00*, Seattle, Washington, USA, 2000.

[FER 00] FERRET O., GRAU B., « A Topic Segmentation of Texts based on Semantic Domains » *Actes de ECAI 2000*, Berlin, Allemagne, 2000, p. 426-430.

[FER 98a] FERRET O., « How to thematically segment texts by using lexical cohesion? », *Actes de ACL-COLING'98 (Student Session)*, Montréal, Canada, 1998, p. 1481-1483.

[FER 98b] FERRET O., ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage, Thèse de doctorat, Université Paris XI, Orsay, 1998.

[FER 98c] FERRET O., GRAU B., « A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts », *Actes de ECAI'98*, Brighton, UK, 1998, p. 155-159.

[FER 98d] FERRET O., GRAU B., MASSON N., « Thematic segmentation of texts: two methods for two kinds of texts », *Actes de ACL-COLING'98*, Montréal, Canada, 1998, p. 392-396.

[FIS 99] FISCUS J., DODDINGTON G., GAROFOLO J., MARTIN A., « NIST's 1998 Topic Detection and Tracking Evaluation (TDT2) », *Actes de DARPA Broadcast News Workshop*, Herndon, Virginia, USA, 1999.

[GRA 84] GRAU B., « Stalking Coherence in the Topical Jungle », *Actes de Fifth Generation Computer System*, Tokyo, Japon, 1984, p. 652-659.

[GRO 86] GROSZ B.J., SIDNER C.L., « Attention, Intentions and the Structure of Discourse », *Computational Linguistics*, vol. 12, 1986, p. 175-204.

[HAL 76] HALLIDAY M.A. K., HASAN R., *Cohesion in English*, Longman, London, 1976.

[HEA 97] HEARST M.A., « TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages », *Computational Linguistics*, vol. 23, n°1, 1997, p. 33-64.

[KAU 99] KAUFMANN S., « Cohesion and Collocation: Using context vector in text segmentation », *Actes de 37th Annual Meeting of the Association for Computational Linguistics (Student Session)*, College Park, USA, 1999, p. 591-595.

[KOZ 93] KOZIMA H., « Text Segmentation Based on Similarity between Words », *Actes de 31th Annual Meeting of the ACL (Student Session)*, Colombus, Ohio, USA, 1993, p. 286-288.

[LIN 97] LIN C.-Y, « Robust Automated Topic Identification », Doctoral Dissertation, University of Southern California, 1997.

[MAS 95] MASSON N., « An Automatic Method for Document Structuring », *Actes de 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, 1995, p. 372-373.

[MO 91] MORRIS J., HIRST G., « Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text », *Computational Linguistics*, vol. 17, n°1, 1991, p. 21-48.

[NOM 94] NOMOTO T., NITTA Y., « A Grammatico-Statistical Approach To Discourse Partitioning », *Actes de 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japon, 1994, p. 1145-1150.

[PAS 97] PASSONNEAU R.J., LITMAN D.J., « Discourse Segmentation by Human and Automated Means », *Computational Linguistics*, vol. 23, n°1, 1997, p. 103-139.

[PIC 00] PICHON R., SÉBILLOT P., « From corpus to lexicon: from contexts to semantic features », dans *PALC'99: Practical Applications in Language Corpora*, *Lodz studies in Language*, vol. 1, Barbara Lewandowska-Tomaszczyk and Patrick James Melia (Eds), Peter Lang, 2000.

[RAS 95] RASTIER F., « La sémantique des thèmes ou le voyage sentimental », dans *L'analyse thématique des données textuelles*, F. Rastier (Eds), Paris, Didier, 1995, p. 223-249.

[REY 94] REYNAR J.C., « An Automatic Method of Finding Topic Boundaries », Actes de *32th Annual Meeting of the Association for Computational Linguistics (Student Session)*, Las Cruces, New Mexico, USA, 1994, p. 331-333.

[SAL 96] SALTON G., SINGHAL A., BUCKLEY C., MITRA M., « Automatic Text Decomposition Using Text Segments and Text Themes », *Actes de Hypertext'96*, Washington, D.C, USA, 1996, p. 53-65.

[SIL 93] SILBERZTEIN M., *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, 1993.

[STE 95] STEIN A. SCHMID H., « Étiquetage Morphologique de Textes Français avec un Arbre de Décisions », *Traitement Automatique des Langues*, vol. 36, n°1-2, 1995, p. 23-35.

[YOU 91] YOUMANS G., « A New Tool for Discourse Analysis: The Vocabulary-Management Profile », *Language*, vol. 67, n°4, 1991, p. 763-789.