Ajustement des frontières de segments thématiques détectés automatiquement

Martine Hurault-Plantet*, Michèle Jardino*
Jean-Baptiste Berthelin*

* LIMSI-CNRS, BP 133, 91403 ORSAY Cedex
{ Martine.Hurault-Plantet, Michele.Jardino, Jean-Baptiste.Berthelin}@limsi.fr
http://www.limsi.fr/

Résumé. La segmentation thématique de textes utilise la cohésion de ces textes, qui peut être établie de façon lexicale ou par l'observation de marqueurs spécifiques. Nous avons, d'une part, exploité la cohésion lexicale des corpus politique, juridique et scientifique de DEFT'06 en leur adaptant un segmenteur probabiliste (Utiyama et Isahara 2001) ; d'autre part, nous avons ajusté les frontières ainsi obtenues, en exploitant des marqueurs de cohésion et de rupture de cohésion appris sur les corpus d'entraînement. Pour chaque corpus, ou pour deux parties d'un corpus hétérogène, un pas d'itération a été choisi pour le segmenteur probabiliste. Les marqueurs linguistiques ont été appris en appliquant aux trois corpus lemmatisés un modèle à n-grammes de mots, ce qui a servi à corriger les frontières obtenues par le premier segmenteur, et à en découvrir d'autres, qu'il n'avait pas détectées, notamment dans le corpus des discours politiques.

1 Introduction

La segmentation thématique de textes se propose pour but de découper un texte en segments consécutifs qui présentent chacun une unité thématique, et qui sont chacun en rupture thématique avec le segment précédant et avec le segment suivant. Son utilisation dans la segmentation de transcriptions de l'oral, et en particulier des journaux télévisés (Stokes *et al.*, 2002), et ses applications en résumé de texte (Farzindar *et al.*, 2004) ou en recherche d'information (Hearst et Plaunt, 1993), en font un domaine de recherche très actuel du traitement du langage.

La plupart des travaux sur le sujet sont sous-tendus par les notions de cohérence et de cohésion d'un texte. D'après Charolles (1995), la propriété de cohérence est caractérisée par l'ensemble des liens sémantiques et pragmatiques qui existent entre les énoncés du texte, alors que la propriété de cohésion est caractérisée par un système de marques dans le texte, telles que les connecteurs et les anaphores, ou encore les expressions introductrices de cadres de discours. Ces marques linguistiques tendent à faciliter l'interprétation des liens de cohérence et à assurer la continuité du texte. La cohérence reflète donc la structure profonde du texte, les inférences crédibles que l'on peut faire à partir des énoncés, alors que la cohésion reflète la structure de surface, la fluidité du discours.

Les théories sur la structure du discours caractérisent la cohérence d'un texte par les relations rhétoriques (Mann et Thompson, 1988) ou intentionnelles (Grosz et Sidner, 1986), ou d'autres relations sémantiques (Polanyi, 1988) qui existent entre les différents segments du discours. Mais pour détecter ces relations marquant la structure profonde du texte, des

études ont été menées sur les corrélations existant entre ces dernières et des marqueurs linguistiques de cohésion, principalement connecteurs et anaphores (Passonneau et Litman, 1993 ; Knott *et al.*, 2000). La cohésion d'un texte a été étudiée sous différents angles : d'une part la cohésion lexicale qui s'appuie sur les répétitions de mots et des relations sémantiques telles que la synonymie ou l'hyperonymie (Morris et Hirst, 1991), et d'autre part les marques de cohésion ou de rupture de cohésion qui s'appuient sur des expressions linguistiques, des connecteurs, ou des anaphores (Piérard *et al.*, 2004).

Depuis le début des années 90, ces différentes analyses ont été utilisées pour développer des segmenteurs qui sont évalués suivant des méthodes et des métriques tendant à se normaliser (Pevzner et Hearst, 2002 ; Sitbon et Bellot, 2004). Une première méthode consiste à faire appel à des juges humains pour segmenter le corpus de test. Mais cela pose des problèmes d'accord entre juges et de taille du corpus à segmenter. Une autre méthode consiste à sélectionner des segments dans des textes venant de différentes sources et de les concaténer. Même si le corpus ainsi constitué est artificiel, il présente une segmentation assez peu contestable et permet de constituer de gros corpus. Mais s'il reflète assez bien les coupures thématiques des journaux télévisés par exemple, en revanche il fournit peu d'éléments sur les problèmes qui se posent dans la segmentation thématique d'un texte d'un seul tenant.

Beaucoup de segmenteurs s'appuient sur la notion de cohésion lexicale (Hearst, 1997 ; Reynar, 1998 ; Choi, 2000). Ils utilisent la répétition de termes soit en calculant des scores de similarité entre phrases, soit en calculant des densités de termes dans l'espace des termes du texte. L'algorithme de segmentation proposé par Utiyama et Isahara (2001) utilise également la répétition de termes, mais pour calculer la probabilité la plus vraisemblable de segmentation à partir d'un modèle statistique du texte. L'avantage de ces segmenteurs est qu'ils sont indépendants du domaine et ne nécessitent pas d'apprentissage. Les marqueurs linguistiques de cohésion ou de rupture de cohésion sont moins largement utilisés que la cohésion lexicale car d'une part ils nécessitent un apprentissage sur de larges corpus, et d'autre part ils sont dépendants du domaine. Néanmoins, ils présentent l'avantage de marquer de façon précise une frontière de segment, alors que les segmenteurs basés sur la cohésion lexicale ont tendance à constituer des regroupements cohérents mais dont les frontières sont moins précises. Ces constatations nous ont conduit à combiner les deux approches, cohésion lexicale et marqueurs linguistiques, pour résoudre la tâche de segmentation thématique proposée par DEFT'06. On peut trouver une approche mixte similaire à la nôtre dans Beeferman *et al.* (1999).

La tâche proposée par DEFT'06 consistait à trouver la première phrase de chaque segment dans trois corpus différents dont la segmentation manuelle d'origine a été conservée. Nous avons d'abord appliqué le segmenteur probabiliste de Utiyama et Isahara (2001) pour obtenir une première segmentation. Nous avons ensuite corrigé les frontières des segments obtenus à l'aide des marqueurs linguistiques de rupture de cohésion appris sur le corpus d'entraînement par un modèle n-grammes de mots. Nous avons enfin fusionné cette segmentation avec une nouvelle segmentation obtenue en sélectionnant comme début de segment toutes les phrases du corpus dont la probabilité de segmentation, calculée sur les marqueurs de rupture de cohésion, était supérieure à un seuil donné.

Après une étude des corpus, nous décrirons le segmenteur probabiliste de Utiyama et Isahara (2001), puis nous présenterons les méthodes d'apprentissage des marqueurs linguistiques. Nous décrirons ensuite les deux utilisations que nous avons faites de ces marqueurs : d'une part la méthode nous permettant de corriger les limites des segments

obtenus par le segmenteur probabiliste, et d'autre part la segmentation obtenue par un seuil sur la probabilité qu'une phrase soit un début de segment. Nous conclurons sur les résultats obtenus sur les corpus de test.

2 Présentation des corpus

Les données sont constituées de trois corpus de genres différents : des discours politiques, des textes juridiques et un ouvrage scientifique. La segmentation de référence des corpus consiste dans les paragraphes des textes d'origines pour les discours politiques, les lois dans le corpus des lois, chaque loi pouvant comporter plusieurs articles, et enfin les sections et sous-sections pour l'ouvrage scientifique.

Les corpus ont été répartis pour 60% en données d'apprentissage incluant une segmentation et 40 % en données de test. Le format des données est d'une phrase par ligne. Le tableau 1 suivant rassemble quelques statistiques sur le découpage en segments du corpus d'apprentissage selon le genre des données.

Genre	Nombre de phrases	Nombre de segments	Nombre moyen de phrases par segment
Discours	303 373	18 929	16
Lois	433 456	9 934	44
Scientifique	4722	337	14

TAB 1 – Statistiques sur la segmentation du corpus d'apprentissage

On constate une grande disparité de tailles entre les trois genres aussi bien pour le nombre de phrases que pour le nombre moyen de phrases par segment. Le corpus scientifique est beaucoup plus petit que les deux autres corpus et le corpus des lois, qui a une taille comparable à celle des discours, a des segments comprenant beaucoup plus de phrases.

Le tableau 2 suivant récapitule le dénombrement des corpus en termes de mots incluant la ponctuation après un pré-formatage consistant à transformer toutes les majuscules en minuscules et les chiffres en un symbole unique.

Genre	Taille du vocabulaire	Nombre total de mots	Nombre moyen de mots par phrase	Nombre moyen de mots par segment
Discours	62 465	8 186 044	27	432
Lois	57 763	11 555 852	27	1163
Scientifique	7147	154 735	33	459

TAB 2 – Statistiques sur les mots du corpus d'apprentissage

On retrouve la différence de taille importante entre le corpus scientifique et les deux autres corpus. Mais les phrases de ces deux derniers sont plus courtes. Les textes de lois ont une particularité : ils comportent des phrases entières dans les différentes langues de la communauté européenne.

En termes de marqueurs d'amorçage d'un segment, nous avons dans un premier temps dénombré le nombre de mots différents qui débutent chaque segment, voir le tableau 3.

Discours (870 mots)	Lois (1 mot)	Scientifique (84 mots)
monsieur (3111)	Article (9934)	l' (38)
je (1205)		la (34)
vous (725)		nous (25)
la (707)		les (25)
mesdames (611)		le (le)
le (581)		dans (16)
il (518)		on (14)
j' (475)		un (12)
mais (453)		il (10)

TAB 3 – Liste des mots initiaux de chaque segment classés en ordre décroissant de fréquence

Pour le corpus des lois le premier mot est toujours *article*. Pour les deux autres corpus ce sont essentiellement des mots outils avec en sus *monsieur*, *mesdames*, *messieurs* pour les discours. Cela nous a incité à rechercher des marqueurs plus avant dans le texte (voir section 4).

3 Le segmenteur probabiliste

3.1 Présentation de l'algorithme

Le segmenteur probabiliste proposé par Utiyama et Isahara (2001) repose sur un modèle statistique du texte à segmenter. Le texte est considéré comme une suite de mots statistiquement indépendants $W = w_1 w_2 ....w_n$. Une segmentation d'un texte W est considérée comme une suite de segments statistiquement indépendants $S = S_1 S_2 ....S_m$. La probabilité d'une segmentation S étant donné un texte W est donnée par :

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)}$$

La segmentation la plus vraisemblable est alors :

$$\hat{S} = \arg\max_{S} P(W \mid S)P(S)$$

La probabilité d'un texte étant donnée une segmentation est re-définie en fonction de la probabilité d'un mot étant donné un segment :

$$P(W \mid S) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} P(w_j^i \mid S_i) \qquad \text{avec :} \qquad P(w_j^i \mid S_i) \equiv \frac{f_i(w_j^i) + 1}{n_i + k} \text{ (loi de Laplace)}$$

où $f_i(w_j^i)$ est le nombre d'occurrences du mot w_j dans le segment i, n_i est le nombre total de mots dans le segment i, et k est le nombre de mots différents dans le segment i.

La probabilité a priori de segmentation *P(S)* est définie en fonction du nombre *n* de mots et du nombre *m* de segments dans le texte *W* :

$$P(S) \equiv n^{-m}$$

Un algorithme de programmation dynamique permet ensuite de trouver la segmentation optimale.[1]

L'une des propriétés du segmenteur décrit est que le nombre de segments obtenus est assez stable et augmente faiblement en fonction de la taille du corpus. Le nombre de segments obtenus est donc proportionnellement plus faible pour un texte long que pour un texte court. Cela est favorisé par la formule utilisée pour calculer la probabilité a priori de segmentation *P(S)*, qui baisse fortement quand le nombre de segments augmente. Cette propriété est intéressante lorsque la segmentation est un préalable au résumé automatique, car un résumé doit prendre en compte les thèmes les plus génériques du texte donné. Mais dans le cas où l'on veut un nombre assez important de segments, il est nécessaire d'adapter l'algorithme.

3.2 Utilisation du segmenteur probabiliste sur les corpus de DEFT

Plusieurs tests sur le segmenteur nous ont montré que les meilleures performances sont obtenues en lemmatisant le corpus, et en sélectionnant les substantifs, adjectifs, noms propres et abréviations. Nous avons donc lemmatisé les corpus avec le Tree-tagger (Schmid, 1999), et sélectionné les mots sur leur catégorie grammaticale. Pour la première partie du corpus des discours politiques qui est en majuscules, la lemmatisation a été effectuée sur une version du texte mis en minuscules, car la lemmatisation sur la version en majuscules donnait un trop grand nombre de faux noms propres.

L'algorithme tel qu'il est implémenté est confronté à un problème de mémoire dès que le texte atteint une longueur supérieure à 1 000 phrases. Par ailleurs, la segmentation de référence des corpus d'apprentissage comporte un grand nombre de segments, environ 2 à 10 pour 100 phrases, et nous avons vu au paragraphe précédent que le nombre de segments obtenus par le segmenteur probabiliste est très faiblement dépendant de la longueur du texte. Nous avons donc choisi d'itérer le processus de segmentation du corpus de façon à obtenir l'ordre de grandeur du nombre de segments du corpus d'apprentissage. Nous avons fait plusieurs essais sur les trois corpus pour déterminer la taille du pas d'itération en fonction du nombre de segments à obtenir. La proportion de segments étant différente pour chacun des corpus, nous leur avons appliqué des pas différents. Le corpus des discours politiques a été séparé en deux parties sur le critère de la casse des lettres pour effectuer sa lemmatisation. Mais la longueur moyenne d'un segment étant sensiblement différente sur les deux parties, nous avons décidé de laisser séparées ces deux parties dans la suite du traitement pour l'apprentissage puis dans le test.

Nous avons retenu les valeurs des pas d'itération du segmenteur qui produisent un encadrement par une valeur plus faible et une valeur plus forte du nombre de segments à obtenir. Pour évaluer le segmenteur, nous avons pris le F-score[2] utilisé pour DEFT'06, calculé en fonction des mesures de rappel et de précision.

[1] Un programme implémentant cet algorithme est disponible à l'adresse http://www2.nict.go.jp/x/x161/members/mutiyama/software.html

[2] Nous n'avons pas utilisé le F-score souple qui prévoit une fenêtre autour du point de segmentation.

$$rappel = \frac{nombre\ de\ débuts\ de\ segment\ corrects\ retrouvés}{nombre\ total\ de\ débuts\ de\ segment\ corrects}$$

$$précision = \frac{nombre\ de\ débuts\ de\ segment\ corrects\ retrouvés}{nombre\ total\ de\ débuts\ de\ segment\ retrouvés}$$

$$Fscore(\beta) = \frac{(\beta^2 + 1) \times précision \times rappel}{\beta^2 \times précision + rappel} \qquad \text{avec } \beta = 1$$

Nous avons constaté que, pour la valeur plus forte du nombre de segments trouvés, on obtient un rappel supérieur à la précision, et pour la valeur plus faible, on obtient une précision supérieure au rappel. Le tableau 4 rassemble les résultats des expérimentations du segmenteur probabiliste sur les corpus d'apprentissage. Pour le corpus de seconde partie des discours (casse avec lettres en minuscules) et pour le corpus des lois qui sont de grande taille, respectivement 269 103 lignes et 433 456 lignes, nous avons pris des extraits.

Corpus d'apprentissage	Discours (partie 1)		Discours (extrait partie 2)		Lois (extrait)		Ouvrage scientifique	
Nombre de phrases	34 270		89 361		5 084		4 722	
Nombre de segments	3 822		5 035		136		337	
Longueur moyenne d'un segment	9		18		37		14	
Pas du segmenteur (en nombre de phrases)	100	50	200	100	500	300	200	100
Nombre de segments trouvés	2 861	4 510	4 506	7 057	120	148	207	371
Rappel	0.25	0.32	0.30	0.38	0.25	0.26	0.23	0.30
Précision	0.33	0.27	0.34	0.27	0.28	0.24	0.38	0.27
F-score	0.29	0.30	0.32	0.32	0.26	0.25	0.29	0.28

TAB 4 – Les résultats des expérimentations du segmenteur probabiliste sur les corpus d'apprentissage.

Nous avons supposé que les corpus de test auraient la même proportion de segments que les corpus d'entraînement correspondants, ce qui nous permet de calculer un nombre approximatif de segments pour chaque corpus de test. Nous avons appliqué aux corpus de test les pas d'itération que nous avons retenus sur les corpus d'entraînement (voir tableau 4), et nous avons constaté que les nombres de segments obtenus sur les corpus de test pour les deux valeurs du pas d'itération encadraient bien la valeur supposée du nombre de segments pour chaque corpus. Le tableau 5 donne les valeurs obtenues sur les corpus de test.

Les scores obtenus sur les corpus d'entraînement nous ont semblés faibles comparés aux résultats décrits dans Utiyama et Isahara (2001). Mais d'une part les corpus que ces auteurs ont utilisés ont été constitués artificiellement de segments pris dans différents textes, ce qui radicalise les ruptures thématiques entre segments. Et d'autre part, le critère d'évaluation adopté est le critère de Beeferman et al. (1999) qui calcule une probabilité d'erreur sur la segmentation et tient compte d'une distance entre une frontière de segmentation trouvée et la frontière de référence.

Corpus de test	Discours (partie 1)		Discours (partie 2)		Lois		Ouvrage scientifique	
Nombre de phrases	20 505		175 891		271 914		3 219	
Nombre supposé de segments	2 280		9 770		6 180		230	
Pas du segmenteur (en nombre de phrases)	100	50	200	100	500	300	200	100
Nombre de segments trouvés	1 679	2 685	8 353	13 735	5 488	7 820	136	236

TAB 5 – Les caractéristiques des corpus de test et les pas d'itération appliqués.

Le problème étant de trouver les limites exactes des segments thématiques, nous avons étudié la possibilité d'utiliser pour cela des marqueurs linguistiques de cohésion et de rupture de cohésion. Nous décrivons dans la section suivante la méthode utilisée pour trouver les marqueurs discriminants, et leur utilisation dans la détermination des limites des segments thématiques.

4 Les marqueurs linguistiques

4.1 Apprentissage des marqueurs linguistiques sur les corpus

Nous avons utilisé un modèle de langage n-grammes de mots pour détecter des marqueurs linguistiques de cohésion et de rupture de cohésion. Ces modèles sont généralement employés dans les systèmes de reconnaissance de la parole (Jelinek, 1998). Ce sont des modèles probabilistes qui prédisent un mot connaissant les *n-1* mots précédents.

Le but que nous voulions atteindre était d'avoir, pour chaque n-grammes de mots situé en début de phrase et en fin de phrase, une probabilité pour que ce n-grammes de mots ouvre ou ferme un segment thématique, ou au contraire soit à l'intérieur d'un segment. Nous devons donc calculer les probabilités conditionnelles d'avoir un point de segmentation entre deux phrases consécutives *A* et *B* étant donné un n-grammes de mots respectivement en début ou en fin de la phrase *A,* ou de la phrase *B*. Pour obtenir ces probabilités, nous avons introduit dans les corpus d'entraînement une marque *SEGMENTATION* à tous les changements de segment et une marque *SUITE* à tous les changements de phrase situés à l'intérieur d'un segment. Les probabilités conditionnelles de segmentation, étant donné le n-grammes de mots $mot_1 \ldots mot_n$, sont obtenues par une simple estimation du maximum de vraisemblance :

$$P(SEGMENTATION \mid mot_1 \ldots mot_n) = \frac{nombre\_occurrences(SEGMENTATION, mot_1 \ldots mot_n)}{nombre\_occurrences(mot_1 \ldots mot_n)}$$

Suivant que le n-grammes de mots $mot_1 \ldots mot_n$ est situé en début ou en fin de la phrase *A* ou de la phrase *B*, on obtient 4 probabilités conditionnelles différentes : la probabilité qu'une phrase soit un début de segment étant donné respectivement, un n-grammes de mots en début de phrase, un n-grammes de mots en fin de phrase, un n-grammes de mots en début de la phrase précédente, un n-grammes de mots en fin de la phrase précédente. La probabilité conditionnelle pour les deux phrases *A* et *B* d'appartenir à un même segment est la

complémentaire à *1* de la probabilité conditionnelle de segmentation dans la même configuration de mots. Ces probabilités ont été calculées sur une partie du corpus d'entraînement puis appliquées sur l'autre partie. Nous nous sommes limités aux unigrammes, bigrammes et trigrammes de mots.

Les résultats obtenus diffèrent suivant les corpus. Le corpus des discours politiques est le seul qui présente plusieurs marqueurs de rupture de cohésion discriminants, c'est-à-dire qui donnent une probabilité de segmentation suffisamment élevée. Le corpus des lois n'en possède pas, cependant tous ses segments commencent par l'expression introductive d'un article de loi *Article X*, que nous avons donc utilisée comme marqueur unique. En revanche, tous les corpus possèdent des marqueurs de cohésion discriminants. Dans les sections suivantes, nous présentons des exemples des marqueurs linguistiques que nous avons trouvés pour le corpus des discours politiques et le corpus scientifique.

4.2 Les marqueurs linguistiques dans les corpus

4.2.1 Le corpus des discours politiques

Les corpus ont d'abord été lemmatisés, et nous prenons en compte tous les lemmes des phrases, y compris la ponctuation. Si le lemme d'un mot n'a pas pu être trouvé par le lemmatiseur, nous prenons le mot. Un marqueur est donc constitué d'une suite de un, deux, ou trois lemmes.

Dans la mesure où le nombre de phrases à l'intérieur d'un segment est largement supérieur au nombre de phrases de début de segment, nous avons considéré qu'un marqueur de rupture de cohésion est discriminant s'il a une probabilité supérieure à *0.5*, c'est-à-dire s'il se trouve plus d'une fois sur deux dans une phrase de début de segment. En revanche un marqueur de cohésion est discriminant si sa probabilité est égale à *1*, c'est-à-dire s'il ne se trouve jamais en début de segment.

Unigrammes		Bigrammes		Trigrammes	
Marqueur	Probabilité	Marqueur	Probabilité	Marqueur	Probabilité
veuillez	1	veuillez_,	1	vif_le_France	1
vif	0.51	bonsoir_.	0.8	veuillez_,_je	1
		merci_,	0.75	je_vous_prier	0.97
		je_leve	0.67	merci_,_monsieur	0.83
		le_democratie	0.57	je_leve_mon	0.67
		vif_le	0.57	et_si_nous	0.67
				vous_voir_donc	0.6
				vous_pouvoir_compter	0.6
				nous_avoir_pouvoir	0.6
				mais_,_le	0.6
				je_avoir_pouvoir	0.6
				avoir_ce_fin	0.6
				donc_,_si	0.57

TAB 6 – Les marqueurs de rupture de cohésion pour les fins de segments.

Nous donnons dans les tableaux 6 et 7 des exemples de marqueurs de rupture de cohésion discriminants, pour la première partie du corpus des discours dont le texte est en majuscules. Les n-grammes de mots en fin de phrase ne nous ont pas donné de marqueurs discriminants, en revanche les n-grammes de début de phrase nous en ont fourni un certain nombre. Le tableau 6 donne les n-grammes de début de phrase qui ont une probabilité supérieure à 50 % d'apparaître dans la phrase précédant le début de segment. Le tableau 7 donne les n-grammes de début de phrase qui ont une probabilité supérieure à 50 % d'apparaître dans une phrase de début de segment.

Unigrammes		Bigrammes		Trigrammes	
Marqueur	Probabilité	Marqueur	Probabilité	Marqueur	Probabilité
bonjour	1	,_le	1	mon_cheres_francaises	1
,	0.87	mon_deuxieme	0.89	avoir_le_invitation	1
sire	0.85	sire_,	0.86	monsieur_le_ambassadeur	0.93
:	0.77	:_le	0.86	monsieur_le_premier	0.9
question	0.71	mon_cheres	0.85	m_._marquer	0.86
monsieur	0.69	mon_troisieme	0.80	je_en_venir	0.78
francaises	0.6	bonsoir_madame	0.73	mon_cher_compatriotes	0.78
troisieme	0.58	monsieur_le	0.70	m_._le	0.76
monseigneur	0.57	Monsieur_le	0.67	bonsoir_madame_,	0.73
bonsoir	0.55	parallelement_,	0.6	monsieur_le_president	0.72
Monsieur	0.53	mon_cher	0.59	m_._bortoli	0.71
		une_question	0.57	monsieur_le_maire	0.68
		et_d'abord	0.57	je_ajouter_enfin	0.67
		madame_le	0.57	Monsieur_le_président	0.64
		avant_de	0.56	enfin_,_vous	0.64
		le_troisieme	0.51	monsieur_le_secretaire	0.62
				je_vouloir_egalement	0.62
				il_falloir_enfin	0.62
				messieurs_le_presidents	0.6
				je_vouloir_enfin	0.6
				monsieur_le_gouverneur	0.6
				monsieur_le_directeur	0.6
				madame_,_mademoiselle	0.6
				avoir_cote_de+le	0.6
				je_souhaiter_egalement	0.6
				le_troisieme_objectif	0.6

TAB 7 – Les marqueurs de rupture de cohésion pour les débuts de segments.

Cette partie du corpus était très majoritairement en majuscules que nous avons donc transformé en minuscules avant la lemmatisation. Les mots ne comportaient donc pas d'accentuation, ce qui augmente les probabilités d'erreur de lemmatisation. Ainsi l'expression *à cette fin* a été transformée par le lemmatiseur en *avoir_ce_fin*. En revanche le lemmatiseur remet parfois des majuscules où il n'y en avait pas (*France*), et des accents sur les verbes trouvés (*être*).

Les marqueurs obtenus caractérisent les débuts de phrase de fin de discours politique. On peut noter la présence de *je_leve_mon* qui est certainement suivi de *verre*.

On peut constater que les introducteurs de discours politique sont beaucoup plus nombreux que les expressions de clôture de discours. Là encore on remarque les effets de la lemmatisation sur un texte sans accents avec *avoir_cote_de+le* pour *à côté du*. Les expressions *bonsoir* comme *bonjour* semblent davantage caractériser les ouvertures de discours que les fermetures, sauf en ce qui concerne *bonsoir* suivi d'un point.

4.2.2 Le corpus scientifique

Dans ce corpus, aucun n-grammes ne caractérise particulièrement les débuts et les fins de segments. Les probabilités de segmentation obtenues sont toutes inférieures à 0.5, il n'y a aucun marqueur de rupture de cohésion discriminant. En revanche on peut caractériser les marques de cohésion par les n-grammes qui ne sont jamais présents dans un début de segment, et qui ont donc une probabilité égale à 1 de se trouver à l'intérieur d'un segment (voir tableau 8).

n-grammes	Marqueurs de cohésion
unigrammes	Par ; légende ; mais ; elle ; cela ; plus ; figure ; chaque ; cependant ;, comme ; son ; soit ; tout ; comment ; finalement ; entrée ; rappeler ; noter ; et ; ensuite ; ceci
digrammes	ce_être ; par_exemple ; il_être ; le_premier ; Legende_le ; il_falloir ; le_méthode ; en_effet ; Legende_un ; nous_aller ; on_avoir ; le_automate ; un_autre ; nous_présenter ; le_fonction ; de_plus ; cependant_, ; si_on ; par_contre ; par_conséquent ; le_risque ; le_figure ; le_exemple ; pour_le ; par_ailleurs ; on_noter ; le_second ; le_ensemble ; finalement_, ; on_y ; on_chercher ; nous_renvoyer ; il_ne ; en_particulier ; de_ce ; un_automate ; le_réponse ; le_langage ; le_deuxième ; il_y, ensuite_, ; elle_être ; de_même ; ce_ne ; ce_algorithme
trigrammes	par_exemple_, ; en_effet_, ; le_méthode_de ; ce_être_pourquoi ; de_plus_, ; dans_ce_cas ; par_contre_, ; le_figure_Reference ; ce_être_le ; par_conséquent_, ; par_ailleurs_, ; le_risque_empirique ; un_exemple_de ; Legende_le_automate ; ce_être_ce ; nous_renvoyer_le ; le_ensemble_de ; il_y_avoir ; en_particulier_, ; ce_ne_être ;

TAB 8 – Marqueurs de cohésion discriminants en début de phrase.

On trouve dans ces listes à la fois les articulateurs logiques décrits par S.Durand[3], tels que *Par exemple*, *Par contre*, *Par ailleurs*, *Finalement*, *De plus*, *En effet*, *Ensuite*, *C'est pourquoi*, mais aussi des déterminants et pronoms anaphoriques comme *Son*, *Elle*, *Ceci*, *Cela*, et des expressions qu'on retrouve fréquemment dans la littérature scientifique comme *Nous allons*, *Nous présentons*, *On note*, *On cherche*, *Nous renvoyons …*

[3] Les articulateurs logiques, cours en ligne de Français Langue Étrangère à l'ENPC, http://www.enpc.fr/fr/formations/depts/dfl/section_fle/ressources/ecrit/articulateurs.htm

5 Combinaison des marqueurs linguistiques et du segmenteur probabiliste

5.1 Correction de la segmentation obtenue par le segmenteur probabiliste

Pour corriger les frontières des segments obtenus par le segmenteur probabiliste, nous recherchons, dans une fenêtre de deux phrases avant et après une frontière, la phrase qui a une probabilité de segmentation supérieure à 0.5. Pour cela nous faisons passer chaque frontière de segment par un arbre de décision basé sur les marqueurs linguistiques.

L'algorithme considère prioritairement la phrase frontière, puis les deux phrases adjacentes et enfin les deux phrases les plus éloignées. Pour chaque phrase, l'algorithme considère d'abord la probabilité qu'elle soit un début de segment, et ensuite la probabilité que la phrase précédente soit une fermeture de segment, en fonction des plus longs marqueurs qu'elle contient. Enfin, si aucune phrase proche ne satisfait les critères c'est la phrase frontière initiale qui est gardée si toutefois sa probabilité de segmentation n'est pas inférieure à un seuil que nous avons fixé à 0.05. Nous voulons ainsi éviter de garder une limite qui a une trop forte probabilité d'être à l'intérieur d'un segment.

Cet algorithme a été utilisé pour corriger les limites des segments du corpus des discours qui contenait un nombre important de marqueurs linguistiques. Pour corriger les limites du corpus des lois, nous avons utilisé le même principe mais en élargissant la fenêtre autour de la phrase frontière. Nous avons simplement recherché dans les dix phrases les plus proches de la phrase frontière, la phrase la plus proche qui contenait l'introducteur *Article X*. Quant au corpus scientifique, nous avons pris la démarche inverse : nous avons enlevé de la liste des débuts de segment obtenue par le segmenteur probabiliste toutes les phrases qui contenaient un marqueur de cohésion de probabilité 1 et dont, par conséquent, la probabilité de segmentation est égale à 0.

5.2 Segmentation par les marqueurs linguistiques

Nous avons constaté que si les corrections effectuées donnaient un score plus élevé que celui obtenu uniquement par le segmenteur, en revanche certains débuts de segments qui contenaient un marqueur de rupture de cohésion discriminant n'étaient pas trouvés par le segmenteur probabiliste.

Nous avons donc décidé de construire une nouvelle segmentation en sélectionnant les phrases dont la probabilité de segmentation était supérieure à un seuil donné. Après plusieurs essais, nous avons retenu le seuil de 0.20 qui donne le meilleur rapport entre la précision et le rappel.

Ce processus de segmentation a été appliqué au corpus des discours politiques et au corpus scientifique. Nous avons, pour ces deux corpus, effectué ensuite une fusion simple des deux segmentations, celle obtenue par la correction des frontières produites par le segmenteur probabiliste et celle obtenue directement par les marqueurs linguistiques.

5.3 Résultats sur les corpus d'entraînement

Nous présentons dans le tableau 9 les résultats que nous avons obtenus sur le corpus d'entraînement. Pour calculer les probabilités de segmentations des marqueurs linguistiques, nous avons pris classiquement une partie du corpus d'entraînement pour l'apprentissage, et l'autre partie pour le test. Les résultats sont conformes à la spécificité des marqueurs observés sur les différents corpus. Dans l'ouvrage scientifique, nous n'avons pas trouvé de marqueurs de rupture de cohésion qui soient discriminants, et les marqueurs de cohésion se révèlent insuffisants pour isoler efficacement les limites des segments. On n'observe qu'une très légère amélioration par rapport au segmenteur probabiliste qui s'appuie sur la cohésion lexicale. L'amélioration est plus nette sur le corpus des discours politiques. La performance du segmenteur basé uniquement sur les marqueurs de rupture de cohésion se révèle légèrement plus efficace que le segmenteur probabiliste. Et la fusion des deux augmente plus sensiblement le F-score. Mais c'est sur le corpus des lois que l'apport d'un marqueur fiable se révèle le plus performant pour corriger les frontières de la segmentation basée sur la cohésion lexicale.

Corpus	Discours politiques			Lois			Ouvrage scientifique		
Rappel, Précision, F-score	R	P	F	R	P	F	R	P	F
Segmenteur probabiliste (cohésion lexicale)	0.38	0.27	0.32	0.26	0.24	0.25	0.30	0.27	0.28
Correction des frontières du segmenteur probabiliste	0.31	0.47	0.38	0.60	0.59	0.59	0.27	0.33	0.30
Segmentation par les marqueurs linguistiques	0.33	0.36	0.35				0.11	0.23	0.14
Fusion	0.46	0.35	0.40				0.35	0.29	0.32

TAB 9 – Résultats des différentes méthodes de segmentation sur les corpus.

6 Résultats du test

Les résultats que nous avons obtenus sur le corpus de test sont assez stables par rapport à ceux que nous avons obtenus sur le corpus d'entraînement. Nous avons fourni trois résultats d'exécution de nos processus de segmentation sur les corpus de test. Les deux premiers ont été obtenus avec la chaîne complète telle que nous l'avons décrite dans la section 5, en prenant pour le segmenteur probabiliste les pas d'itération retenus en section 3.2. Ces pas sont de 50 et 100 phrases pour la première partie du corpus des discours, 100 et 200 phrases pour la seconde partie du corpus des discours et pour le corpus scientifique, et 300 et 500 phrases pour le corpus des lois. Pour la troisième exécution, nous avons utilisé le segmenteur probabiliste tout seul. Le tableau 10 montre les résultats obtenus sur le corpus de test. Les organisateurs de DEFT'06 ont calculé trois F-score différents correspondant à la formule donnée en section 3.2, un F-score strict qui considère le point de segmentation exact comme seul correct, et deux F-score souples qui prennent en compte une fenêtre de une (F-score souple 1) ou deux phrases (F-score souple 2) autour du point de segmentation.

Corpus	Discours politiques			Lois			Ouvrage scientifique		
F-score strict, souple 1, souple 2	strict	souple 1	souple 2	strict	souple 1	souple 2	strict	souple 1	souple 2
Fusion avec un pas minimum	0.371	0.463	0.535	0.531	0.537	0.599	0.279	0.379	0.459
Fusion avec un pas maximum	0.373	0.460	0.524	0.524	0.524	0.579	0.283	0.367	0.445
Segmenteur probabiliste	0.237	0.421	0.547	0.173	0.398	0.493	0.305	0.442	0.554

TAB 10 – Résultats sur le corpus de test.

Le segmenteur probabiliste donne des résultats assez semblables sur les discours politiques et sur l'ouvrage scientifique, et un peu moins bons sur le corpus des lois. En revanche, conformément à nos essais sur le corpus d'apprentissage, c'est le corpus des lois qui est le mieux corrigé. On remarque que le segmenteur probabiliste double son score entre le F-score strict et le F-score souple avec une fenêtre de deux phrases. Il repère donc mieux les segments eux-mêmes que leurs frontières.

Nos résultats se comparent favorablement aux moyennes des participations à DEFT'06, qui sont, pour le F-score strict, de 0.181 pour le corpus des discours, 0.170 pour le corpus des lois, et 0.115 pour le corpus scientifique.

Pour estimer la difficulté de la tâche, nous avons réalisé une segmentation manuelle des données sur des échantillons constitués par les 1000 premières phrases de chacun des trois corpus d'entraînement[4]. Le tableau 11 en donne les résultats en termes de F-score en détection stricte et en détection à plus ou moins 2 phrases.

Genre	Discours	Lois	Scientifique
Détection stricte	0.46	0.56	0.47
Détection à ± 2 phrases	0.63	0.56	0.54

TAB 11 – F-score des segmentations manuelles d'échantillons de1000 phrases des trois corpus d'entraînement.

On remarque que la segmentation manuelle donne des résultats similaires à la segmentation automatique sur le corpus des lois et on observe également pour les deux types de segmentation que la détection souple apporte très peu d'amélioration. Pour les deux autres corpus, discours et scientifique, la segmentation manuelle est plus performante en détection stricte mais redevient analogue à la segmentation automatique en détection souple.

[4] Pour une meilleure représentativité des échantillons, nous aurions pu en choisir plusieurs sur tout le corpus plutôt qu'un seul en début de corpus, mais la segmentation manuelle est une tâche assez lourde et notre objectif était d'avoir un ordre de grandeur pour pouvoir comparer la tâche manuelle et la tâche automatique.

7 Conclusion

Notre système de segmentation de textes, qui combine l'action d'un segmenteur probabiliste et l'utilisation de marqueurs de rupture de cohésion, a donné des résultats supérieurs à la moyenne de ceux obtenus par l'ensemble des participants de DEFT'06. Globalement nous obtenons des F-scores qui vont d'environ 30% à 50% suivant les corpus. La comparaison avec une évaluation manuelle sur un ensemble réduit montre la difficulté de la tâche puisque également des F-scores d'environ 50% sont alors obtenus. Cette difficulté est due en partie au niveau de granularité de la tâche. Séparer des textes de provenances différentes est peut-être plus simple que de retrouver des paragraphes ou des sections dans un texte. Filippova et Strube (2006) qui ont étudié plus spécifiquement l'identification des paragraphes d'un texte, estiment que si une rupture thématique tombe en général sur une limite de paragraphe, l'inverse n'est pas toujours vrai.

Néanmoins, plusieurs voies d'amélioration sont à envisager. Pour le segmenteur probabiliste, comme l'indique Ituyama et Isahara (2001), d'autres expressions de la probabilité a priori de la segmentation pourraient être utilisées. Pour les marqueurs de ruptures thématiques, nous remarquons que les plus efficaces que nous ayons trouvé sont des marques de début et de fin de discours. Les marqueurs de rupture thématique intra-textuelle sont en revanche plus difficiles à trouver, comme le montre la différence entre les résultats obtenus sur les discours politiques (voir section 4.2.1) et ceux obtenus sur l'ouvrage scientifique (voir section 4.2.2). C'est donc sur les marqueurs intra-textuels qu'il nous faudra progresser en nous référant aux études sur la structure organisationnelle du discours et les marqueurs permettant de la déterminer (Charolles, 1997).

Références

Beeferman D., Berger A. et Lafferty J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3), pp.177-210.

Charolles M. (1995). Cohésion, cohérence et pertinence du discours, *Travaux de Linguistique,* 29, pp. 125-150.

Charolles M. (1997). L'encadrement du discours, univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, LANDISCO, URA-CNRS 1035, Université Nancy2, 6, pp. 1-73.

Choi F.Y.Y. (2000). Advances in domain independent linear text segmentation. *Actes de NAACL'00*, Seattle, USA, pp. 26-33.

Filippova K. et Strube M. (2006). Using linguistically motivated features for paragraph Boundary identification. Actes de EMNLP 2006, Sydney, Australia, pp. 267-274.

Farzindar A., Lapalme G. et Desclés J-P. (2004). Résumé de textes juridiques par identification de leur structure thématique. *TAL*, 45(1), pp.1-26.

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

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

Hearst M. A. et Plaunt C. (1993). Subtopic structuring for full document access. *Actes de SIGIR'93*, Pittsburgh, PA, USA, pp. 59-68.

Jelinek F. (1998). Statistical Methods for Speech Recognition. MIT Press.

Knott A., Oberlander J., O'Donnell M., Mellish C., (2000). Beyond elaboration: the interaction of relations and focus in coherent text. In T.Sanders, J.Schilperood et W.Spooren (eds.). *Text representation : linguistic and psycholinguistic aspect.* Amsterdam : Benjamins, pp. 181-196.

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

Passonneau R.J. et Litman D.J. (1993). Intention-based segmentation : human reliability and correlation with linguistic cues. *Actes de ACL'93*, Columbus, Ohio**,** pp. 148-155.

Pevzner L. et Hearst M.A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1), pp.19-36.

Piérard, S., Degand, L., & Bestgen Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, Mars 2004, pp. 859-864

Polanyi L. (1988). A formal model of discourse structure. *Journal of Pragmatics*, 12, pp. 601-638.

Reynar J.C. (1998). Topic segmentation: algorithms and applications. *Ph.D. thesis*, University of Pennsylvania.

Schmid H. (1999). Improvements in Part-of-Speech Tagging with an Application To German. In Armstrong, S., Chuch, K. W., Isabelle, P., Tzoukermann, E. & Yarowski, D. (Eds.), *Natural Language Processing Using Very Large Corpora*. Dordrecht : Kluwer Academic Publisher.

Sitbon, L. et P. Bellot (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. *Actes de TALN 2004*. Fès, Maroc.

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. *Actes de ACL'01*, Toulouse, France, pp. 491-498.

Summary

Text segmentation by topic relies upon the cohesion of given texts, derived either from lexical measurements or from specific markers. We addressed both kinds of cohesion in the DEFT'06 political, juridical and scientific corpora. Lexical cohesion was exploited by adjusting a probabilistic segmenter (Utiyama and Isahara 2001) to these texts, and explicit cohesion was explored by means of positive and negative cohesion markers that were learned from training corpora. For each corpus, or for different parts on a heterogeneous one, we determined an iteration step for the probabilistic segmenter. Linguistic markers were learned by applying a word n-gram model to the three lemmatized corpora. This gave us a means for correcting the boundaries found by the first segmenter, and for finding some quite new ones, particularly in the political corpus.