TALN 2010, Montréal, 19–23 juillet 2010

Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique

Clémentine Adam [1] Philippe Muller [2] Cécile Fabre [1]
(1) CLLE / Université de Toulouse
(2) IRIT / Université de Toulouse & Alpage / INRIA
adam@univ-tlse2.fr, muller@irit.fr, cfabre@univ-tlse2.fr

Résumé. Cette étude a pour but de contribuer à la définition des objectifs de la segmentation thématique (ST), en incitant à prendre en considération le paramètre du type de textes dans cette tâche. Notre hypothèse est que, si la ST est certes pertinente pour traiter certains textes dont l'organisation est bien thématique, elle n'est pas adaptée à la prise en compte d'autres modes d'organisation (temporelle, rhétorique), et ne peut pas être appliquée sans précaution à des textes tout-venants. En comparant les performances d'un système de ST sur deux corpus, à organisation thématique "forte" et "faible", nous montrons que cette tâche est effectivement sensible à la nature des textes.

Abstract. This paper aims to contribute to a better definition of the requirements of the text segmentation task, by stressing the need for taking into account the types of texts that can be appropriately considered. Our hypothesis is that while TS is indeed relevant to analyse texts with a thematic organisation, this task is ill-fitted to deal with other modes of text organisation (temporal, rhetorical, etc.). By comparing the performance of a TS system on two corpora, with either a "strong" or a "weak" thematic organisation, we show that TS is sensitive to text types.

Mots-clés : Segmentation thématique, organisation textuelle, cohésion lexicale, voisins distributionnels.

Keywords: Text segmentation, textual organisation, lexical cohesion, distributional neighbours.

1 Introduction

La tâche de segmentation thématique (ST), qui consiste à délimiter des segments textuels de contenu homogène sur la base d'indices de rupture lexicale, a fait la preuve de sa faisabilité et de son apport dans différentes tâches de TAL (Hearst, 1997; Chen *et al.*, 2009). Même si certains travaux ont cherché à prendre en compte des indices de rupture (*cue-phrases*) de différentes natures – par exemple (Litman & Passonneau, 1995) –, cette tâche repose généralement sur l'hypothèse que les textes s'organisent principalement selon un plan thématique, chaque thème se singularisant par le recours à un vocabulaire suffisamment spécifique pour le distinguer des autres. Cette hypothèse est pourtant loin de faire consensus dans les travaux menés sur le discours, qui montrent au contraire que l'organisation thématique n'est qu'un mode d'organisation des textes parmi d'autres (Péry-Woodley & Scott, 2006). Elle n'est pas pertinente pour tous les types de textes, et elle n'est pas exclusive, pour un même texte, d'autres types d'organisation alternatifs. En particulier, de nombreux travaux ont étudié les textes sous l'angle de leur organisation rhétorique, qui

s'articule autour de segments identifiés comme des unités fonctionnelles déterminées par des buts argumentatifs spécifiques, 'argumentative moves' (Swales, 1990), 'argumentative zoning' (Teufel, 1999). Ces études ont montré que différents segments sont caractérisables par des faisceaux de traits linguistiques de nature essentiellement grammaticale, et ne considèrent pas forcément la répartition du vocabulaire comme un critère discriminant (Biber *et al.*, 2007). De fait, il ne va pas du tout de soi que l'on puisse aborder par les mêmes méthodes de segmentation des textes organisés thématiquement, rhétoriquement, temporellement, voire par une combinaison de ces modes, et que les indices lexicaux soient toujours discriminants pour placer des ruptures entre segments textuels. Les tableaux 1 et 2 donnent des exemples de textes tirés de Wikipédia de manière à illustrer cette diversité des modes d'organisation. La liste des titres de section de premier niveau donne un bon aperçu de la façon dont le texte s'organise. Le tableau 1 montre des textes dont l'organisation thématique est manifeste.

Le Malawi	Le panda géant
- Histoire	- Historique
- Politique	- Légende
- Géographie	- Alimentation
- Économie	- Reproduction
- Démographie	- Protection
- Culture	

TABLE 1 – Exemples d'organisation textuelle thématique

Le tableau 2 montre d'abord un exemple d'organisation temporelle, typique des biographies, qui se clôt par une partie bilan. Les deux autres textes (leadership et mythe) illustrent un mode de progression rhétorique (sur le principe des 'moves' de Swales) qui permet dans ces deux cas d'organiser la présentation d'une notion selon un schéma argumentatif similaire : d'abord définir la notion, puis présenter une typologie, enfin détailler certaines de ses instances.

Laurent Truguet	Leadership	Mythe
- Jeunesse jusqu'à la Rév.	- Terminologie	- Définition
- sous la Révolution	- Types de leadership	- Aspects du mythe
- L'Empire	- Caractéristiques du leadership	- Typologie et éléments du mythe
- sous la Monarchie	- Le leadership de droit et de fait	- Postérité du mythe
- Le bilan	- Le paradigme des leaderships multiples	

TABLE 2 – Exemples d'organisation textuelle non thématique

L'impact des types de textes sur la procédure de ST a rarement été pris en considération par les travaux qui mettent en oeuvre cette tâche – exception faite de (Ferret *et al.*, 1998) –, au point que, comme le déplorent (Bestgen & Piérard, 2006), les mêmes algorithmes sont parfois appliqués à une tâche de segmentation de texte et de délimitation de textes concaténés. Les expériences de ST sont généralement menées sur des types de textes qui se prêtent intuitivement à cette approche – par exemple les articles encyclopédiques sur les villes chez (Chen *et al.*, 2009) ou (Adam & Morlane-Hondère, 2009) –, sans qu'on cherche à établir explicitement la nature des textes qui sont adaptés à la tâche.

Notre objectif est d'intégrer le paramètre du type de textes dans la tâche de segmentation en comparant les performances d'un système de ST sur deux groupes de textes, déterminés selon leur propension à s'organiser plutôt thématiquement ou à obéir à d'autres principes de présentation - rhétorique, temporelle. Nous montrons dans cet article que la tâche de ST est effectivement sensible à la nature des textes, en

montrant que même une approche relativement naïve de la notion de type de texte permet de faire émerger des différences significatives de performances. Après avoir brossé un panorama des méthodes actuelles en ST (section 2), nous décrivons le système de ST que nous avons mis en oeuvre pour cette étude (section 3) ; nous présentons ensuite l'expérimentation qui vise à comparer ses performances selon les deux types de textes (section 4), et discutons les résultats obtenus (section 5).

2 La segmentation thématique : un panorama des méthodes

La segmentation thématique a pour but le découpage linéaire d'un texte en unités présentant une cohérence autour d'un sujet. La grande majorité des approches de ST se fondent sur la méthode initiée par (Hearst, 1997) : le texte est divisé en blocs contigus correspondant à une unité fixée à l'avance (un nombre n de mots, de phrases ou de paragraphes), puis on définit une fenêtre glissante qui parcourt le texte linéairement et permet de calculer un score de similarité à chaque intersection entre blocs dans le texte. Une méthode de segmentation s'attache alors à trouver les points où la similarité présente des évolutions fortes, interprétées comme des indications de rupture de la continuité thématique. Une alternative à cette approche par pavage (*tiling*) est de supposer des thèmes sous-jacents qu'il s'agit de détecter : chaque unité du texte considéré est rapportée à un ou plusieurs thèmes, et la segmentation consiste à trouver ces thèmes (Chen *et al.*, 2009; Ferret, 2007). Les thèmes peuvent être prédits par des « topic models » (Chen *et al.*, 2009), une forme d'Allocation de Dirichlet Latente (ADL), qui sont associés à des distributions lexicales différentes, ou bien par des associations lexicales calculées à partir des textes, par exemple par un *clustering* en amont (Ferret, 2007). Une fois les thèmes identifiés pour chaque unité de texte, les segments correspondent aux blocs d'unités contiguës partageant le même thème.

Dans les approches par pavage, la mesure de similarité entre blocs la plus simple est basée sur le nombre de répétitions lexicales (souvent uniquement les noms), rapporté au nombre des unités présentes.Les variantes consistent alors à jouer sur le lissage de l'évolution de la similarité en prenant en compte des contextes différents (plus de blocs voisins, et plus d'interactions entre eux) ou sur la normalisation des liens de cohésion lexicale, avec des mesures de *tf.idf* locales par exemple (Malioutov & Barzilay, 2006). Mais la similarité peut également dériver d'autres sources : collocations (Ferret, 2002), similarité dans un espace lexical de dimension réduite par exemple par analyse sémantique latente (Choi *et al.*, 2001; Bestgen & Piérard, 2006) – proche de l'ADL mentionnée ci-dessus –, ou bien similarité de distribution des unités lexicales (Adam & Morlane-Hondère, 2009), toutes méthodes qui sont censées apporter une forme de lissage pour prévenir d'éventuels fossés dans les répétitions de forme, dans la mesure où elles font émerger une large gamme de liens de proximité sémantique.

Dans le système que nous avons développé et que nous décrivons dans la section suivante, nous comparons deux types de similarité, en utilisant d'une part les répétitions simples et d'autre part une mesure de similarité distributionnelle. Nous aurions pu faire appel à d'autres approches utilisant des similarités plus riches, mais pour notre propos (la comparaison des performances d'un système de ST en fonction des types de textes), il est suffisant que notre système ait des performances comparables à celles de l'état de l'art. Par ailleurs les approches génératives (*topics models*) ont l'inconvénient de devoir être entraînées sur des textes représentatifs des thèmes à retrouver, ce qui rend la technique un peu circulaire dans le cadre de ce que nous cherchons à étudier [1].

1. Nous pourrons de toute façon étendre la portée de la comparaison dans une étude plus large, notamment en transposant l'approche de Ferret au réseau lexical induit par la similarité distributionnelle.

3 Description du système de segmentation thématique mis en oeuvre

Notre système de ST, développé dans le cadre du projet VOILADIS [2], utilise une approche linéaire, à la manière de (Hearst, 1997), et fait appel pour le calcul des scores de similarité lexicale à une base de voisins distributionnels. La base de voisins distributionnels utilisée a été générée à partir d'un corpus constitué de l'ensemble des articles de la version francophone de Wikipédia, soit plus de 470000 articles pour 194 millions de mots. Le programme d'analyse distributionnelle est lancé en aval des sorties de l'analyse SYNTEX (Bourigault, 2007). Les triplets syntaxiques <gouverneur, relation, dépendant> (ex : <départ,POUR,destination>) fournissent les données permettant de rapprocher les paires de voisins en utilisant la mesure de Lin.

Nous donnons ci-dessous le détail de notre chaîne de traitement :

– Les liens de voisinage, éventuellement pondérés par leur score de Lin, sont projetés sur les textes ; les répétitions sont également prises en compte, et dotées d'un score de 1. Quelques paramètres de filtrage des voisins sont testés : seuils sur le nombre de voisins que peut avoir un mot, et sur l'écart quadratique moyen de ces voisins à la position du mot (ainsi, les mots ayant peu de voisins ou des voisins proches de leur position dans le texte seront favorisés)
– Le texte est parcouru par une fenêtre glissante, afin de calculer localement des scores de cohésion. L'unité de segmentation, ainsi que la taille de la fenêtre en nombre d'unités, sont paramétrables. Les unités de segmentation possibles sont : (i) la phrase ; (ii) le bloc de mots de taille fixe. Par exemple, si l'unité choisie est la phrase, et que la taille de la fenêtre est fixée à 6, on calculera à la fin de chaque phrase un score basé sur le nombre de liens entretenus par le groupe de trois phrases qui précède, et le groupe de trois phrases qui suit (fig. 1) ; ce nombre est normalisé par le nombre de liens possibles (le produit des noms, verbes et adjectifs situés à gauche et à droite de la fenêtre).

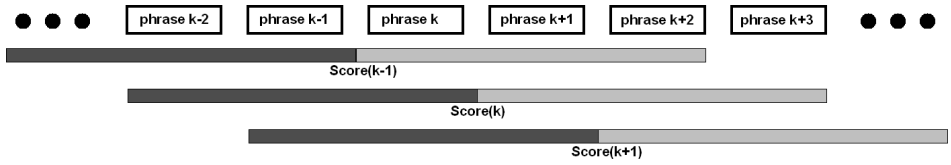

FIGURE 1 – Représentation de la fenêtre glissante avec -fen=3 et -unit=phrase

– La courbe des scores obtenue est lissée ; nous avons opté pour un lissage gaussien [3], avec deux paramètres ajustables : le nombre d'itérations et le degré du lissage. La figure 2 présente les courbes brute et lissée pour l'article Wikipédia *Bulgarie*. Les barres verticales indiquent la segmentation de référence (c'est-à-dire les positions des titres de section).
– Les vallées (creux de la courbe) dont la profondeur dépasse un écart-type à la moyenne des profondeurs sont considérées comme correspondant aux ruptures du texte. Ces ruptures, qui, selon l'unité choisie, se trouvent dans le meilleur des cas à la frontière d'une phrase, mais peuvent également intervenir en plein milieu d'une phrase, sont ramenées à la frontière de paragraphe la plus proche, ce qui produit le texte segmenté final.

On constate que de nombreux paramètres sont ajustables dans notre système ; nous les récapitulons dans le tableau 3. C'est pourquoi dans la phase expérimentale de notre étude, qui fait l'objet de la prochaine section, nous recourons à un corpus de développement pour optimiser ces paramètres.

2. Projet financé par le PRES de Toulouse
3. En fait une estimation par noyau, le noyau étant gaussien ; il correspond ici à une moyenne sur le voisinage de chaque

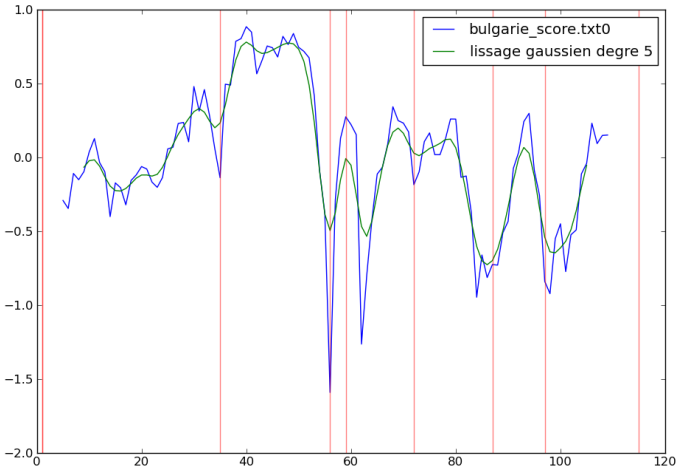

FIGURE 2 – Exemple de courbe avec lissage

Paramètre	Description	Valeurs
-unit	unité de segmentation	phrase / bloc
-bloc	taille du bloc	<nb mots>
-fen	taille de la demi-fenêtre glissante	<nb blocs>
-it	nb d'itérations du lissage	<nb>
-deg	degré du lissage	<nb>
-lin	pondération des liens par score de Lin	oui / non
-filtNb	seuil sur le nb max de voisins différents par item	non ou <nb voisins>
-filtPos	seuil sur l'écart moyen des voisins à la position de l'item	non ou <nb tokens>

TABLE 3 – Paramètres de notre système de ST

4 Procédure et évaluation

L'hypothèse que nous souhaitons valider par cette expérience est que le recours à la ST se justifie pour des textes dont la structuration est effectivement ressentie comme thématique, mais n'est pas motivé pour aborder d'autres modes d'organisation textuelle. Ainsi, nous voulons inciter à mieux définir quel peut être l'objet de la tâche de ST, et à ne pas appliquer cette tâche sans précaution à des textes tout-venant.

Caractérisation du corpus utilisé Nous avons pour cette expérience constitué deux sous-corpus à partir de la version d'avril 2007 de l'encyclopédie en ligne Wikipédia (nous avons choisi de sélectionner des textes qui appartiennent tous au corpus sur lequel la ressource lexicale a été construite). Les articles ont été extraits de manière automatique, sur la base de critères de sélection fixés par nous. Les critères de sélection sont les suivants : pour être retenu, un article doit avoir au minimum 1000 mots, au moins 4 titres de sections (qui fournissent la segmentation de référence), et un maximum de 2 niveaux de profondeur de titres (une profondeur trop importante aurait amené à faire des choix délicats quant aux titres retenus pour la segmentation de référence) ; il doit également appartenir à une liste de catégories établie. Nous avons en effet pris le niveau des catégories définies dans l'encyclopédie comme critère de répartition des textes dans les deux sous-corpus. Le sous-corpus à organisation thématique forte (corpus THEM) rassemble des textes consacrés à la description de pays, de villes et d'animaux, dont on sait qu'ils se prêtent généralement

point, pondérée selon la distance au point selon une gaussienne.

bien à une organisation thématique. Le sous-corpus à organisation thématique faible (corpus NON-THEM) réunit des biographies, dont l'organisation est typiquement temporelle, et des textes présentant des notions abstraites, des concepts, pour lesquels nous avons montré (tab. 2) que l'approche thématique est généralement mal adaptée. L'intervention humaine se concentre donc en amont de la constitution du corpus, par la définition des critères de sélection et de répartition dans les sous-corpus. Aucun traitement n'est effectué en aval (post-sélection, nettoyage, etc.). La caractérisation des corpus obtenus est donnée dans le tableau 4. Nous précisons pour chaque sous-corpus le nombre de paragraphes, qui correspond au nombre de segments potentiels pour notre système de ST et le nombre de titres de premier niveau, c'est-à-dire le nombre de ruptures dans notre segmentation de référence.

	nb textes	nb mots	nb para.	nb seg. (titre niv=1)	nb seg/txt	nb seg/para
corpus THEM	344	578346	10953	3051	8,869	0,279
corpus NON-THEM	210	387941	6454	1182	5,629	0,183

TABLE 4 – Caractérisation des corpus THEM et NON-THEM

Corpus de développement et optimisation des paramètres La procédure de segmentation décrite section 3 dépend de nombreux paramètres dont les conséquences ne sont pas toujours prédictibles *a priori*. Beaucoup d'auteurs fixent des paramètres similaires selon des critères empiriques pas toujours explicites. Nous avons choisi d'isoler une partie du corpus de départ pour l'utiliser comme un corpus de développement explicite, sur lequel nous avons fait varier un certain nombre de paramètres afin d'ajuster la segmentation. Pour cela, nous avons extrait au hasard un peu moins de 10% du corpus rassemblé initialement, en prenant autant (*i.e.* 21) de textes des sous-corpus THEM et NON-THEM (rappelons que le corpus initial n'est pas tout à fait équilibré ; nous avons équilibré celui de développement pour ne pas favoriser la classe majoritaire). Les variations faites sur les 8 paramètres sujet de l'optimisation ont généré plus de 2000 configurations.Nous avons conservé la configuration ayant obtenu les meilleurs résultats selon l'indice classique WindowDiff (noté WD) de comparaison de segmentation ; elle est donnée dans le tableau 5.

-unit	-bloc	-fen	-it	-deg	-lin	-filtNb	-filtPos
bloc	10 mots	10 blocs	2	3	non	10 voisins max.	500 tokens

TABLE 5 – Configuration de paramètres retenue

Évaluation Pour évaluer les résultats de la segmentation, nous prenons comme référence les positions des titres de premier niveau au sein des articles ; pour comparer les résultats du système de ST à cette référence, nous appliquons les mesures classiques pour cette tâche : les indices Pk et WindowDiff. Ces mesures sont moins strictes sur les positions des bornes de segments que la précision et le rappel, qui ne permettent pas de juger de la proximité d'une prédiction avec la borne réelle. Les deux mesures Pk et WD « adoucissent » l'évaluation en estimant le nombre moyen de bornes correctes dans une fenêtre de taille donnée projetée sur le texte. Nous avons ajouté une mesure proposée par (Bestgen, 2009), appelée par lui « distance de Hamming généralisé » et qui est en fait une distance d'édition avec des coefficients particuliers pour les coûts d'insertion/d'effacement/de déplacement, rapportée au nombre de points de coupure possibles. Elle est notée "edit" dans nos tables de résultats. La distance d'édition est censée corriger certains biais de WD, elle-même censée corriger certains biais de Pk ; nous ne rentrons pas dans les détails ici, les mesures étant relativement cohérentes entre elles sur nos résultats [4]. Ces mesures étant

4. On peut se référer à (Georgescul *et al.*, 2006) pour une discussion de la pertinence des procédures d'évaluation de la ST.

souvent difficiles à interpréter et à comparer, la table 6 donne les résultats pour la pire configuration, la configuration moyenne et la meilleure configuration, que nous allons appliquer au reste de notre corpus. Il faut noter que ces mesures sont des mesures de distance, la distance de la référence à elle-même est donc 0., et un score plus bas indique une plus grande proximité avec la référence.

configuration	pk	wd	edit	nb seg/txt
référence	0,000	0,000	0,000	7,67
meilleure	0,294	0,299	1,611	5,27
moyenne	0,321	0,329	1,787	5,18
pire	0,352	0,369	1,983	6,47

TABLE 6 – Résultats sur le corpus de développement avec différentes configurations de paramètres

5 Résultats et analyse

Nous avons appliqué notre système de ST, avec la configuration de paramètres optimisée sur le corpus de développement, sur les deux sous-corpus THEM et NON-THEM. Les tables 7 et 8 synthétisent les résultats. Nous donnons deux résultats pour notre système, selon les types de liens de cohésion lexicale pris en compte : répétitions simples de lemmes, ou voisinage distributionnel. Nous avons en outre généré deux *baselines* simplifiées qui permettent de donner une idée des écarts que l'on peut avoir sur les mesures Pk et WD, qui ne sont pas nécessairement simples à interpréter. La première *baseline* (nommée plus bas « hasard exact ») place des ruptures au hasard, mais en nombre correspondant à la référence. Elle permet de contrôler la facilité de se rapprocher des vraies bornes par rapport au nombre moyen de segments rapporté à la taille du texte. Une deuxième *baseline* (« hasard bruité ») proche consiste à perturber le nombre exact de ruptures, en le faisant varier au hasard dans un intervalle de 30% du vrai nombre de ruptures.

Méthode	Pk	WD	edit	nb seg/txt
référence	0	0	0	7,89
baseline "hasard bruité"	0,3659	0,3738	1,6492	9,46
baseline "hasard exact"	0,3417	0,3452	**1,5789**	7,89
répétitions	0,3114	0,3144	1,5907	4,93
voisins	**0,3091**	**0,3129**	1,5837	5,09

TABLE 7 – Résultats pour le sous-corpus THEM

Méthode	Pk	WD	edit	nb seg/txt
référence	0	0	0	8,07
baseline "hasard bruité"	0,3569	0,3616	1,8032	6,68
baseline "hasard exact"	**0,3149**	**0,3181**	**1,5645**	8,07
répétitions	0,3612	0,3662	1,8846	5,08
voisins	0,3613	0,3676	1,9291	5,16

TABLE 8 – Résultats pour le sous-corpus NON-THEM

Une autre indication de la représentativité des scores pourrait être prise dans la littérature, même si la variété des approches, des entrées et des évaluations (vrai textes ou concaténations artificielles) doit inciter à la prudence. Si l'on se réfère au très récent (Chen *et al.*, 2009), qui opère sur un corpus similaire à une partie du nôtre (les articles de villes dans le Wikipédia anglais), l'état de l'art précédent représenté par

(Eisenstein & Barzilay, 2008) atteint un Pk de 0,317 et un WD de 0,376 sur ce corpus, avec la connaissance du nombre de segments ; l'approche de (Chen *et al.*, 2009) à base de topic models enrichis de contraintes globales atteint quant à elle sur leur meilleure configuration les très bons scores de 0,28 pour le Pk et de 0,25 pour le WD, sans connaissance du nombre de segments, mais en posant une borne supérieure sur le nombre de thèmes présents *dans tout le corpus* (fixée à 10 ou 20 thèmes), ce qui limite un peu la généralisation.

Au vu de nos résultats, on constate que l'hypothèse globale d'une différence entre les deux types de textes THEM et NON-THEM se vérifie assez nettement, quelle que soit la métrique considérée, et que les algorithmes de segmentation choisis sont meilleurs sur les textes du sous-corpus THEM, même si les variances (non rapportées dans le tableau) sont importantes.

Pour évaluer les différences entre méthodes, (Chen *et al.*, 2009) affirment que les tests de significativité statistique sur cette tâche ne sont pas standardisés, et ils n'en reportent pas. Ceux qui font de tels tests utilisent un *t-test* sans préciser s'il est apparié ou pas, (Choi *et al.*, 2001; Galley *et al.*, 2003; Ferret, 2007). Nous avons pour notre part fait le test des rangs signés de Wilcoxon entre les séries de scores des baselines et des méthodes par cohésion, appariées par texte, test qui ne suppose rien sur la distribution *a priori* des scores, au contraire du t-test. Alors qu'on observe des valeurs de $p<0,01$ pour la différence entre les *baselines* et les algorithmes de segmentation sur l'expérience THEM, la différence n'est pas significative pour les textes du sous-corpus NON-THEM (et on constate que les variances sont plus fortes). La *baseline* exacte dans le cas NON-THEM présente des résultats étonnants, dont il faudrait vérifier s'ils ne sont pas directement liés au nombre de prédictions plus élevé que fait cette méthode (il s'agirait alors d'un effet secondaire indésirable des mesures de comparaison de segmentation).

Une hypothèse secondaire de ce travail était que les liens induits par similarité distributionnelle étaient plus informatifs et devaient avoir un impact sur l'évaluation globale. Concernant cet aspect, on ne peut que constater la proximité des scores sur les deux sous-corpus (et le test de Wilcoxon n'indique pas de différence significative).

Étant donnée la forte variance que nous avons observée sur les résultats, y compris sur le sous-corpus THEM, nous avons évalué les résultats par catégories Wikipédia à l'intérieur des corpus THEM et NON-THEM (ces résultats sont récapitulés dans le tableau 9). On constate que les résultats des deux sous-corpus

thème	Pk	WD	edit	nb par./seg.	nb par.	nb textes
animaux	0,2794	0,2803	1,4076	0,1651	25,5724	145
pays	0,3443	0,3472	1,7695	0,1223	40,7353	136
concepts	0,3488	0,3510	1,8377	0,1632	30,4706	68
villes	0,3508	0,3541	1.7610	0,1348	31,1250	32
personnes	0,3738	0,3777	1.9062	0,1778	26,6132	106
autres NON-THEM	0,4041	0,4112	1,7337	0,1655	31,7333	15

TABLE 9 – Résultats par sous-catégories par Pk/WD croissant

se reportent sur les catégories qui les composent, à l'exception de la catégorie *concepts* qui obtient des résultats légèrement meilleurs que ceux de la catégorie *villes*. Encore une fois les variances sont fortes. Il s'avère que notre découpage volontairement grossier *a priori* (dans un souci de ne pas trop biaiser l'étude) pourrait s'affiner – à condition de poser clairement les paramètres de ce que nous avons appelé pour l'instant le caractère thématique fort ou faible des textes –, mais qu'il semble valide.

6 Conclusion et perspectives

Nous avons montré dans cette étude que les types de textes ont un impact important sur la ST, et qu'il s'agit donc d'un paramètre à ne pas négliger dans le cadre de cette tâche. Néanmoins, le bilan de l'expérience menée, s'il comporte la confirmation de l'hypothèse de départ, doit être mitigé par des résultats effectifs sur la tâche. Même si les résultats sont proches de ceux de l'état de l'art sur le corpus THEM (surtout si on tient compte du fait que le nombre de segments n'est jamais donné), ils montrent des variances très fortes sur les textes, et n'ont pu confirmer le rôle d'une similarité lexicale plus riche que la simple répétition de formes. Il ressort de l'observation du corpus que les données que nous avons recueillies étaient finalement assez hétérogènes, avec des sections de longueurs très différentes qui ont posé de gros problèmes aux approches qui se basent sur un niveau de variation moyen. La diversité des niveaux de finalisation des articles de Wikipédia explique en particulier la succession de paragraphes très développés et de paragraphes réduits à une seule phrase.

Concernant la segmentation de référence, la subdivision par sections n'est pas toujours le bon mode de segmentation. On trouve beaucoup de sections hétérogènes sur le plan thématique parce que la répartition thématique est faite au niveau des sous-sections (ex : une section « Domaines influencés par le positivisme » se décline en sous-sections : « médecine », « philosophie », « enseignement », « droit », etc.). L'intérêt de prendre la structuration en titres comme segmentation de référence était de fournir facilement des données annotées, et nous avions bien sûr conscience du bruit que cela devait entraîner par rapport à la tâche évaluée. Cela nous a fourni une première analyse qui nous incite à reprendre ces données pour aller vers une évaluation moins artificielle. Mais au-delà d'un simple nettoyage qui court le risque d'être biaisé par l'objectif, on peut aussi poser le problème autrement et partir de l'observation des endroits où les programmes coupent, pour chercher à déterminer si ces lieux sont « interprétables », plutôt que de chercher un alignement avec une segmentation de référence problématique.

Enfin, la question du mode de différenciation des textes à traiter se pose également. Le fait de choisir de comparer des textes appartenant à un même genre textuel (l'article d'encyclopédie), limite leur diversité. La distinction que nous avons considérée se situe au niveau du sujet traité. C'est un premier point d'entrée, qui n'est pas entièrement satisfaisant. Le rapport que nous avons posé au préalable entre catégorie de sujet et type d'organisation n'est pas systématiquement vérifié : si les articles traitant de personnalités sont quasi systématiquement organisés temporellement, certaines notions sont malgré tout traitées de manière au moins partiellement thématique. Une étude privilégiant cette fois une distinction par genre de textes permettrait d'établir un classement sur des critères plus fiables, et d'opposer des textes au fonctionnement plus marqué, renforçant sans doute le contraste déjà observé.

Références

ADAM C. & MORLANE-HONDÈRE F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. In *Actes du colloque RECITAL, Senlis, France.*

BESTGEN Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de textes ? In *Actes de TALN'09*, Senlis, France.

BESTGEN Y. & PIÉRARD S. (2006). Comment évaluer les algorithmes de segmentation automatiques ? Essai de construction d'un matériel de référence. *Actes de TALN : Verbum ex machina, Louvain-la-neuve*, **6**, 407–414.

BIBER D., CONNOR U. & UPTON T. (2007). *Discourse on the move : Using corpus analysis to describe discourse structure*. John Benjamins Publishing Co.

BOURIGAULT D. (2007). *Un analyseur syntaxique operationnel : Syntex*. CNRS & Université de Toulouse-Le Mirail.

CHEN H., BRANAVAN S., BARZILAY R. & KARGER D. (2009). Content Modeling Using Latent Permutations. *Journal of Artificial Intelligence Research*, **36**, 129–163.

CHOI F. Y. Y., WIEMER-HASTINGS P. & MOORE J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, p. 109–117, Pittsburgh.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *EMNLP '08 : Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 334–343, Morristown, NJ, USA : Association for Computational Linguistics.

FERRET O. (2002). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In *TALN'02 : 9e conférence sur le Traitement Automatique des Langues Naturelles*, p. 155–164, Nancy, France.

FERRET O. (2007). Finding document topics for improving topic segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 480–487, Prague, Czech Republic : Association for Computational Linguistics.

FERRET O., GRAU B. & MASSON N. (1998). Thematic segmentation of texts : two methods for two kinds of texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, p. 392–396 : Association for Computational Linguistics.

GALLEY M., MCKEOWN K. R., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In E. HINRICHS & D. ROTH, Eds., *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, p. 562–569, Sapporo, Japan.

GEORGESCUL M., CLARK A. & ARMSTRONG S. (2006). An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, p. 144–151, Sydney, Australia : Association for Computational Linguistics.

HEARST M. A. (1997). TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

LITMAN D. & PASSONNEAU R. (1995). Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, p. 108–115 : Association for Computational Linguistics.

MALIOUTOV I. & BARZILAY R. (2006). Minimum cut model for spoken lecture segmentation. In *ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 25–32, Morristown, NJ, USA : Association for Computational Linguistics.

PÉRY-WOODLEY M.-P. & SCOTT (2006). Discours et Document : traitements automatiques. Numéro thématique. *revue T.A.L.*, **47(2)**, 7–19.

SWALES J. (1990). *Genre analysis : English in academic and research settings.* New York :Cambridge University Press.

TEUFEL S. (1999). *Argumentative Zoning : Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh.