HAL
open science

Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique

Clémentine Adam, François Morlane-Hondère

To cite this version:

Clémentine Adam, François Morlane-Hondère. Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. RECITAL'09, Jun 2009, Senlis, France. pp.article 6. hal-00417489

HAL Id: hal-00417489

https://hal.science/hal-00417489v1

Submitted on 16 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique

Clémentine Adam François Morlane-Hondère
CLLE / Université de Toulouse & CNRS
adam@univ-tlse2.fr, morlanehondere@gmail.com

Résumé. Cette étude s'insère dans le projet VOILADIS (VOIsinage Lexical pour l'Analyse du DIScours), qui a pour objectif d'exploiter des marques de cohésion lexicale pour mettre au jour des phénomènes discursifs. Notre propos est de montrer la pertinence d'une ressource, construite par l'analyse distributionnelle automatique d'un corpus, pour repérer les liens lexicaux dans les textes. Nous désignons par *voisins* les mots rapprochés par l'analyse distributionnelle sur la base des contextes syntaxiques qu'ils partagent au sein du corpus. Pour évaluer la pertinence de la ressource ainsi créée, nous abordons le problème du repérage des liens lexicaux à travers une application de TAL, la segmentation thématique. Nous discutons l'importance, pour cette tâche, de la ressource lexicale mobilisée ; puis nous présentons la base de voisins distributionnels que nous utilisons ; enfin, nous montrons qu'elle permet, dans un système de segmentation thématique inspiré de (Hearst, 1997), des performances supérieures à celles obtenues avec une ressource traditionnelle.

Abstract. The present work takes place within the Voiladis project (Lexical neighborhood for discourse analysis), whose purpose is to exploit lexical cohesion markers in the study of various discursive phenomena. We want to show the relevance of a distribution-based lexical resource to locate interesting relations between lexical items in a text. We call *neighbors* lexical items that share a significant number of syntactic contexts in a given corpus. In order to evaluate the usefulness of such a resource, we address the task of topical segmentation of text, which generally makes use of some kind of lexical relations. We discuss here the importance of the particular resource used for the task of text segmentation. Using a system inspired by (Hearst, 1997), we show that lexical neighbors provide better results than a classical resource.

Mots-clés : Cohésion lexicale, ressources lexicales, analyse distributionnelle, segmentation thématique.

Keywords: Lexical cohesion, lexical resources, distributional analysis, text segmentation.

1 Introduction

L'étude de la structure du discours est un champ de la linguistique qui a suscité de nombreux travaux depuis les années soixante, et qui bénéficie actuellement d'un regain d'intérêt lié aux enjeux qu'il soulève pour le traitement automatique des langues. L'automatisation de la mise au jour des structures discursives pourrait en effet avoir un impact important sur toute application de TAL nécessitant une vision des textes allant au delà du simple « sac de mots » : fouille de données textuelles, résumé automatique, navigation intra-documentaire, etc. (Péry-Woodley & Scott, 2006).

L'analyse du discours repose sur l'observation selon laquelle un texte n'est pas une simple succession de phrases, mais un tout cohérent. Cette cohérence, propriété intrinsèque des textes, est reflétée par les observables que sont les marques de cohésion. Le concept de cohésion englobe tous les phénomènes qui permettent de relier entre elles les phrases d'un texte, participant ainsi à créer sa *texture* (Halliday & Hasan, 1976). Les procédés cohésifs classiquement considérés, dans la continuité d'Halliday & Hassan, sont la référence, la substitution, l'ellipse, la conjonction et surtout la cohésion lexicale, qui est reconnue comme étant le principal vecteur de texture (Hoey, 1991).

Le projet VOILADIS [1] (VOIsinage Lexical pour l'Analyse du DIScours), dans lequel s'inscrit cette étude, a pour but d'utiliser des indices lexicaux pour la mise au jour de phénomènes discursifs, dans une visée d'automatisation. Ce champ est encore peu exploré : la cohésion lexicale reste peu exploitée sur le plan applicatif car elle est difficile à appréhender, et donc à repérer automatiquement. En effet, elle réside généralement dans des relations *non classiques*, que les lexiques ne recensent pas (Morris & Hirst, 2004).

Dans le cadre du projet VOILADIS, la ressource mobilisée est une base de voisins distributionnels : l'analyse distributionnelle automatique de grands corpus permet en effet de rapprocher des mots présentant des contextes d'apparition similaires, lesquels ont tendance à être liés par une relation sémantique qui va souvent au-delà des classifications traditionnelles. Cette méthode permet également de disposer d'une ressource qui reflète véritablement les relations qui opèrent sur un texte donné, dans le sens où la base distributionnelle est avant tout le fruit de l'analyse syntaxique du corpus. L'objectif à long terme du projet VOILADIS, encore en phase exploratoire, est d'évaluer l'apport d'une telle ressource à différentes approches du discours.

Dans une première étape, nous nous sommes intéressés à une application qui a tout particulièrement exploité les indices de nature lexicale : la segmentation thématique. Cette approche assez empirique du discours vise l'identification de segments textuels, de blocs homogènes du point de vue de leur objet. Cette tâche est parmi les plus tributaires des phénomènes de cohésion, dans le sens où les zones thématiques ne sont définies que par le fait qu'elles se trouvent être particulièrement cohésives, ce qui laisse à penser que plus le repérage des relations lexicales sera efficace, mieux les zones seront définies. Pour évaluer la pertinence de notre base de voisins distributionnels pour le repérage des liens de cohésion lexicale, nous avons donc choisi de mesurer son apport à un système de segmentation automatique.

Dans la suite de cet article, nous discutons de la dépendance de la segmentation thématique à une prise en compte de la cohésion lexicale, qu'elle soit basique ou plus fine. Puis nous décrivons la base de voisins que nous avons mobilisée, et plus largement, la méthode qui a permis de

1. Projet du PRES Toulouse coordonné par Cécile Fabre impliquant des chercheurs des laboratoires IRIT (équipe LiLac) et CLLE-ERSS (axes TAL et S'caladis).

la construire, et discutons *a priori* de sa pertinence pour appréhender des relations lexicales variées. Enfin, nous relatons la démarche que nous avons suivie pour montrer l'apport de cette ressource à un système de segmentation thématique.

2 Segmentation thématique et cohésion lexicale

Le but de la segmentation thématique est d'effectuer le pavage d'un texte en segments consécutifs censés présenter une homogénéité du point de vue de leurs thèmes. Cette tâche peut permettre d'améliorer les performances de diverses applications : recherche d'information – (Callan *et al.*, 1992), entre autres, a montré qu'un système de recherche d'information gagne à indexer des unités inférieures au document –, résumé automatique (Brunn *et al.*, 2001), extraction d'information, etc.

De nombreux algorithmes ont été développés pour la segmentation thématique, que l'on peut *grosso modo* regrouper en deux familles (Hernandez, 2004) : (a) ceux qui parcourent linéairement le texte selon une fenêtre d'observation glissante, et procèdent donc de manière ascendante et (b) ceux qui calculent une matrice de similarité pour l'ensemble des unités du texte avant de décider où placer les ruptures, procédant donc de manière descendante (Malioutov & Barzilay, 2006). Les systèmes les plus connus représentant ces deux familles sont d'une part l'algorithme *TextTiling* de (Hearst, 1997), et d'autre part l'algorithme *C99* de (Choi, 2000).

Malgré la variété des approches, la différence entre algorithmes n'apparaît pas cruciale. Ce qui compte avant tout, ce sont les indices utilisés pour mesurer la similarité entre unités de segmentation, que ce soit au niveau local ou global. Pour mesurer la « force » de la cohésion lexicale entre deux pans de texte, on se base sur le nombre de liens (éventuellement pondérés) qu'entretiennent les unités lexicales qu'ils contiennent. Ces liens peuvent être de natures diverses.

Beaucoup de systèmes de segmentation thématique se cantonnent aux liens de répétition lexicale, c'est-à-dire aux répétitions de formes, de formes tronquées ou de lemmes (Hearst, 1997; Choi, 2000) ; une extension consiste à prendre en compte les répétitions de n-grammes, en leur attribuant un poids plus important (Beeferman *et al.*, 1997). L'inconvénient de ces approches est que les scores sont alors basés sur un nombre très restreint d'occurrences, et que beaucoup de liens participant à la cohésion sont donc ignorés. Pour pallier ce problème, il est nécessaire de faire appel à une ressource extérieure. Cette solution est toujours présentée, à notre connaissance, comme permettant d'améliorer les performances des systèmes, au point que les auteurs mettent parfois plus l'accent sur la ressource utilisée que sur l'originalité de leur algorithme. Les ressources mobilisées varient beaucoup du point de vue des relations lexicales qu'elles permettent de détecter.

Certains utilisent une ressource générique, construite à partir d'un dictionnaire ou d'un thésaurus (Kozima, 1993; Lin *et al.*, 2004; Morris & Hirst, 2004). Ainsi, les liens de synonymie, et dans le meilleur des cas ceux relevant d'autres relations classiques telles que l'hyperonymie et l'antonymie, peuvent être pris en compte. D'autres s'appuient sur des ressources construites en corpus (Choi *et al.*, 2001; Bolshakov & Gelbukh, 2001; Ferret, 2002). Les méthodes de constitution de ces ressources varient, mais tournent toujours autour de l'extraction de collocations ou de cooccurrences. Par exemple, l'analyse sémantique latente (ASL) permet d'évaluer la proximité sémantique entre des couples de mots en fonction de leurs cooccurrences au sein de mêmes phrases, paragraphes ou textes sur l'ensemble d'un corpus (Choi *et al.*, 2001).

Les auteurs prônant des approches basées sur corpus font généralement état de meilleures performances. En effet, les liens lexicaux mis en jeu dans la cohésion lexicale vont bien au delà des relations traditionnelles. Classiquement, suivant (Halliday & Hasan, 1976), on distingue entre :

– **Des relations de réitération**, qui englobent la répétition lexicale, la reprise par un synonyme ou un hyperonyme, voire d'autres relations paradigmatiques classiques telles que l'antonymie et la méronymie ;
– **Des relations dites de collocation**, qui associent des mots présentant une tendance à apparaître ensemble, mais ne relevant pas de la réitération. Il s'agit pas nécessairement de relations d'ordre syntagmatique, l'acception du terme « collocation » étant ici plus vaste. Des études (Morris & Hirst, 2004) ont montré que chez les lecteurs, les relations les plus pertinentes pour le repérage des structures discursives étaient dans la plupart des cas des relations échappant aux typologies traditionnelles. Lorsqu'il s'agit d'interpréter un texte, les relations comme la synonymie, l'antonymie, etc. cèdent le pas à des relations « non classiques », moins facilement définissables car plus dépendantes des mots entre lesquels elles se manifestent (*chien/aboyer*, *abeille/miel*, etc.).

Ainsi, construire une ressource à partir d'un corpus permet d'appréhender des relations plus variées, et plus pertinentes pour la structuration du discours. Nous nous inscrivons dans cette veine. Mais nous ne nous limitons pas à l'extraction de collocations ou de cooccurrences : l'originalité de notre ressource est qu'elle ne se cantonne pas aux relations syntagmatiques. Construite grâce à l'analyse distributionnelle d'un corpus, elle repose sur des informations linguistiques plus riches, susceptibles de mettre au jour des relations d'ordre paradigmatique.

3 Notre ressource : les voisins distributionnels

La base de voisins que nous avons utilisée pour cette étude a été générée par le programme Upéry (Bourigault, 2002) à partir d'un corpus constitué de l'ensemble des articles de la version francophone de Wikipédia, soit plus de 470 000 articles pour 194 millions de mots [2]. Ces données ont été préalablement traitées par l'analyseur syntaxique Syntex (Bourigault, 2007), qui utilise l'analyse en dépendance pour générer une représentation du texte particulièrement propice à la méthode distributionnelle.

Le processus permettant d'obtenir une base de voisins distributionnels à partir d'un corpus se divise en trois étapes :

1. le corpus est étiqueté avec TreeTagger [3] ;
2. il est ensuite traité par Syntex, qui extrait les relations syntaxiques sous forme de triplets <gouverneur, relation, dépendant>. Ainsi, le programme va repérer dans la phrase *Pierre mange un biscuit* les triplets <manger, suj, Pierre> et <manger, obj, biscuit>. Quand la relation de dépendance syntaxique se fait via une préposition, cette dernière prend la place de la relation au sein du triplet (*biscuit au chocolat* se représente <biscuit, à, chocolat>) ;
3. afin de faciliter leur traitement par le logiciel Upéry, les triplets obtenus sont ramenés sous la forme de couples <prédicat, argument> où le prédicat correspond au gouverneur auquel on accole la relation, et où l'argument correspond au dépendant (<biscuit, à, chocolat> devient <biscuit_à, chocolat>). Cette formalisation va permettre d'opérer un double rapprochement : celui des prédicats partageant les mêmes arguments, mais aussi celui des

2. Il s'agit d'une version de Wikipédia datant d'avril 2007. Son traitement ainsi que la création de la base de voisins sont dus au travail de Franck Sajous (CLLE-ERSS).

3. Université de Stuttgart (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).

arguments partageant les mêmes prédicats. Concrètement, dans notre base de voisins, le prédicat *manger_obj* est rapproché de *se nourrir_de* via les arguments *pousse*, *bourgeon*, *crustacé*, etc., et l'argument *biscuit* est rapproché de *sucre* via les prédicats *fabrique_de*, *marque_de*, *production_de*, etc. Lors de cette étape, le programme attribue à chaque paire de voisins un score de proximité qui indique dans quelle mesure les distributions des deux mots rapprochés sont similaires. Ce score est obtenu par la mesure de Lin (Lin, 1998) ; il se déroule en deux étapes. La première consiste à calculer la *quantité d'information* (QI) de chaque prédicat et argument, c'est-à-dire le rapport du nombre d'arguments/prédicats avec lesquels ils se combinent dans le corpus, sur la totalité des arguments/prédicats avec lesquels ils pourraient se combiner. Dans un second temps, il s'agit de rapprocher les arguments/prédicats entre eux en s'appuyant sur le nombre de prédicats/arguments qu'ils partagent et de diviser la QI de ces cooccurrents syntaxiques communs aux deux voisins (multipliée par 2) par la somme des QI respectives de chacun des voisins.

Cette méthode possède l'avantage de permettre les rapprochements intercatégoriels qui font cruellement défaut aux ressources actuellement disponibles. Ainsi, le verbe prédicat *manger_obj* se retrouve également associé à des prédicats nominaux comme *bouillon_de*, *cuisson_de*, *recette_de* ou *plat_de* via des arguments communs comme *viande*, *poulet*, *poisson*, *spaghetti*, etc.

La base ainsi obtenue compte environ quatre millions de couples, qui exhibent des relations très hétérogènes : des études comme (Bourigault & Galy, 2005) ou (Fabre & Bourigault, 2006) ont montré qu'il est difficile de dégager une typologie des relations de voisinage extraites ; c'est la conséquence de l'application des méthodes d'analyse distributionnelle à un corpus non spécialisé, qui présente moins de redondance et donc des restrictions syntaxiques moins fortes. En contrepartie, cette ressource offre la possibilité de capter un large éventail de relations de proximité sémantique, à condition d'introduire des filtres sur les scores de voisinage.

4 Voisins distributionnels et segmentation thématique

Le but de cette expérience est de montrer la pertinence du voisinage distributionnel pour détecter les liens de cohésion lexicale, en nous appuyant sur les résultats d'un système de segmentation thématique. Nous avons à cet effet implémenté un algorithme de segmentation inspiré de *TextTiling* (Hearst, 1997) et basé uniquement sur la prise en compte de liens lexicaux. Nous avons soumis au système développé un même corpus en spécifiant chaque fois des liens différents : uniquement des liens de répétition lexicale dans un premier temps ; des liens de synonymie repérés à partir d'un dictionnaire de synonymes [4] dans un deuxième temps ; et enfin, des liens de voisinage distributionnel. Nous décrivons dans la suite de cette section les différentes étapes de cette expérience : constitution du corpus, projection des liens lexicaux, application de l'algorithme de segmentation thématique et évaluation.

Corpus Le corpus que nous utilisons est constitué de 30 articles issus de l'encyclopédie en ligne Wikipédia (dans la version ayant servi à construire la base de voisins distributionnels). Ces articles traitent tous de lieux – pays (par exemple *Danemark*) ou villes (*Salzbourg*). En effet,

4. Le dictionnaire *Dicosyn*. Développé au CRISCO (Université de Caen), il regroupe les synonymes présents dans sept dictionnaires classiques, à savoir le Bailly, le Benac, le Du Chazaud, le Guizot, le Lafaye, le Larousse et le Robert. Il compte environ 49 000 entrées pour 396 000 relations synonymiques et est consultable en ligne à l'adresse suivante : http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

selon nos observations, dans cette catégorie d'articles, les différentes sections correspondent généralement à différents « thèmes » (histoire, géographie, culture, etc.). Cela nous permet ainsi de justifier l'utilisation des titres comme ruptures de référence lors de l'évaluation de la segmentation effectuée. Le corpus est divisé en 1584 paragraphes (donc $1584 - 30 = 1554$ ruptures possibles [5]) et contient 302 titres de sections (donc 302 ruptures de référence).

Projection des liens cohésifs sur le corpus Nous relions dans le corpus des couples de mots, sans pondérer le lien qu'ils entretiennent. Seuls les liens allant au delà de la phrase sont pris en compte. Pour notre première *baseline*, toutes les répétitions de lemmes de noms, de verbes et d'adjectifs sont indiquées. Pour la seconde, toutes les paires de synonymes recensées par notre dictionnaire sont projetées. Pour projeter les voisins, nous avons dû fixer (de manière empirique) différents seuils dus au caractère pléthorique de la ressource : les couples projetés sont ceux dont le score de *Lin* dépasse 0.25 et pour lesquels chaque membre du couple est parmi les 15 meilleurs voisins de l'autre membre.

Nous proposons dans les figures 1, 2 et 3 des visualisations des liens obtenus avec les trois approches décrites, pour un extrait de l'article *Slovaquie*. On peut constater la rareté des liens

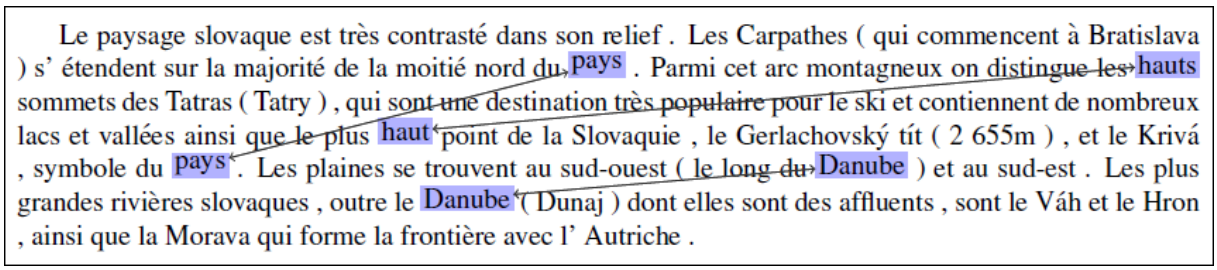

FIGURE 1 – Liens de répétition

de répétition : seulement 3 liens là où la synonymie et le voisinage permettent de détecter respectivement 7 et 8 liens. La répétition est toutefois dans cet exemple la seule méthode qui permet de tisser des liens cohésifs entre noms propres (ici, *Danube*). Les liens de synonymie

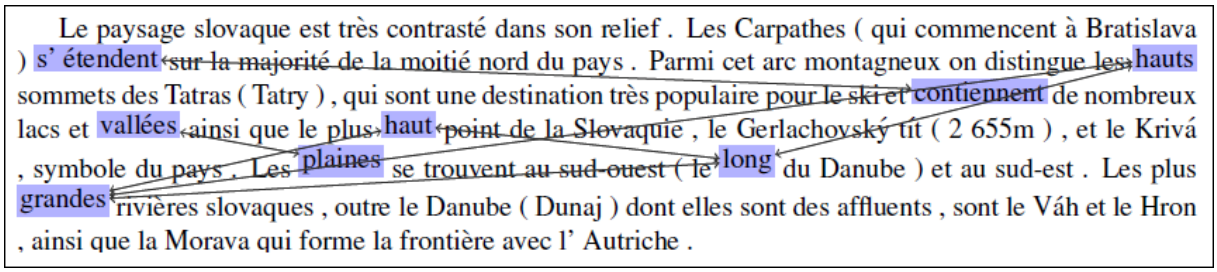

FIGURE 2 – Liens de synonymie

concernent majoritairement des adjectifs, avec *grand*, *haut* et *long*, tous synonymes entre eux. Le lien de synonymie entre *s'étendent* et *contiennent* est difficilement interprétable, en tout cas dans ce contexte : on touche ici aux limites d'une ressource constituée *in abstracto*, dans le sens où les relations lexicales qu'elle recense ne sont en aucun cas adaptées à un type de texte ou à un corpus en particulier, contrairement à la base de voisins qui est construite de façon dynamique. Dans cet exemple, les relations de voisinage permettent de lier des mots répertoriés

5. Pour chacun des 30 textes de n paragraphes, on a $n - 1$ ruptures possibles.

Le paysage slovaque est très contrasté dans son relief . Les Carpathes (qui commencent à Bratislava) s' étendent sur la majorité de la moitié nord du pays . Parmi cet arc montagneux on distingue les hauts sommets des Tatras (Tatry) , qui sont une destination très populaire pour le ski et contiennent de nombreux lacs et vallées ainsi que le plus haut point de la Slovaquie , le Gerlachovský štít (2 655m) , et le Krivá , symbole du pays . Les plaines se trouvent au sud-ouest (le long du Danube) et au sud-est . Les plus grandes rivières slovaques , outre le Danube (Dunaj) dont elles sont des affluents , sont le Váh et le Hron , ainsi que la Morava qui forme la frontière avec l' Autriche .

FIGURE 3 – Liens de voisinage distributionnel

par *Dicosyn* comme étant des synonymes (*plaine/vallée*), des co-hyponymes (*nord/sud-est/sud-ouest*), mais également des mots qui entretiennent des relations moins faciles à catégoriser comme *pays/frontière*, ou *frontière/nord*. Ces dernières relations sont celles qui nous sont le plus précieuses, puisque leur repérage est une des spécificités de notre méthode. Et même si l'on pourrait considérer que le premier couple, *pays/frontière*, relève d'une relation de méronymie, il est difficile de donner un nom à la relation *frontière/nord*. Dans la mesure où ces deux mots font partie du même champ sémantique (celui de la géographie) et que leur mode de liaison échappe à toute classification, on peut considérer qu'on a là un cas de collocation au sens de (Halliday & Hasan, 1976). Indépendamment des performances que nous présenterons dans la suite de l'article, on voit déjà que les voisins présentent un intérêt évident du fait qu'ils mettent au jour des liens qu'aucune ressource classique ne permettrait de détecter.

L'algorithme de segmentation Pour la segmentation des textes du corpus, nous avons opté pour une approche linéaire, par fenêtre glissante, à la manière de (Hearst, 1997). Cette approche n'est pas forcément la plus performante, mais nous l'avons préférée en raison de sa simplicité d'implémentation ; en effet, nous ne poursuivons pas ici l'efficacité, mais la comparaison entre différentes ressources ; l'important pour nous était donc avant tout d'appliquer le même algorithme pour chaque ressource, quel que soit cet algorithme.

Notre unité de base est la phrase ; la fenêtre d'observation que nous appliquons pour calculer les scores de similarité est d'une taille de 6 unités (paramètre conseillé par (Hearst, 1997)). Ainsi, à la fin de chaque phrase, un score basé sur les liens entretenus par deux blocs de trois phrases est calculé (figure 4).

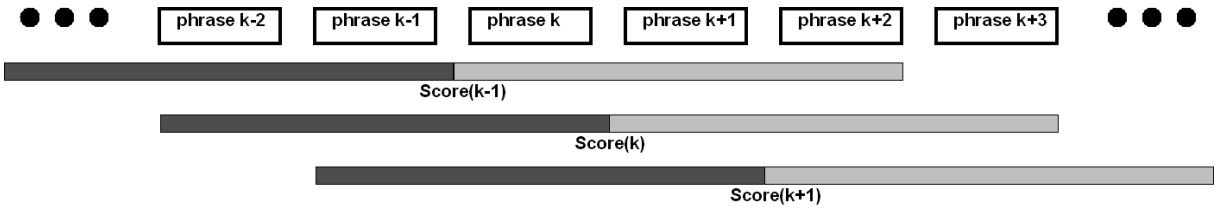

FIGURE 4 – Représentation du calcul du score avec fenêtre glissante

Le score calculé est le suivant : $S = \log\left(\frac{N_{liens}}{N_{liens\ possibles}}\right)$. Le nombre de liens N_{liens} est le nombre de couples de mots jugés similaires à cheval entre les deux blocs de trois phrases. Le nombre de liens possibles $N_{liens\ possibles}$ est le produit des nombres de mots pouvant être liés dans chaque bloc (c'est-à-dire des noms, adjectifs et verbes).

La courbe des scores calculés est ensuite lissée. Toutes les *vallées* sont repérées, et leurs profondeurs calculées. On appelle *vallée* un point de la courbe qui est entouré par des points de

valeurs plus élevées (c'est-à-dire un *minimum local*). Pour déterminer la profondeur d'une vallée, on remonte de part et d'autre du point considéré tant que l'on rencontre des valeurs plus élevées ; la profondeur de la vallée est calculée en faisant la moyenne des deux différences calculées (à gauche et à droite du point). Les vallées dont la profondeur dépasse l'écart-type à la moyenne sont considérées comme correspondant aux ruptures du texte. Ces ruptures, situées entre deux phrases, sont ramenées à la frontière de paragraphe la plus proche, ce qui produit le texte segmenté final.

Décisions prises sur l'évaluation Évaluer un système de segmentation thématique est délicat. De nombreux problèmes sont soulevés, et peuvent *grosso modo* être ramenés à deux questions : (a) Quelle référence ? (b) Quel score d'évaluation ?

(a) Pour évaluer la segmentation automatique, il faut la comparer à une segmentation de référence. Certains font pour cela appel à des annotations manuelles, mais font généralement état d'accords inter-annotateurs très faibles. D'autres prennent le parti d'accoler bout à bout des séquences appartenant à des textes différents ; les ruptures thématiques sont alors les ruptures entre textes. Cette position pose un problème évident de circularité : on fabrique l'objet que l'on postule. Pour cette expérience, nous avons décidé d'utiliser comme ruptures de référence les positions des titres de sections.

(b) Les scores habituels de précision et de rappel ne sont pas adaptés pour évaluer un système de segmentation thématique. En effet, ils ne permettent pas de rendre compte du fait qu'une rupture proche de la rupture de référence est meilleure qu'une rupture éloignée. D'autres scores ont été proposés, dont les plus usités sont les mesures *Pk* (Beeferman *et al.*, 1999) et *WindowDiff* (Pevzner & Hearst, 2002). La mesure *Pk* consiste à compter le nombre de fois où deux mots pris au hasard à une distance k sont dans le même segment à la fois dans la référence et dans l'hypothèse. La mesure *WindowDiff* consiste à calculer la différence du nombre de ruptures dans une fenêtre glissante. Nous donnons ici nos résultats selon ces deux mesures.

Résultats Nous reportons dans le tableau 1 les résultats obtenus par l'algorithme de segmentation appliqué au corpus décrit, selon les liens cohésifs pris en compte (répétition, synonymie ou voisinage distributionnel). Les scores affichés correspondent aux moyennes des scores obtenus pour chaque texte. Il est à noter qu'avec les mesures *Pk* et *WindowDiff*, un score moins élevé reflète de meilleures performances. Pour mettre en perspective les résultats présentés, nous rapportons également les résultats obtenus avec des ruptures placées au hasard, le nombre de ruptures de références étant approximativement [6] connu.

Liens pris en compte	Pk	WindowDiff
Hasard	0.436	0.452
Répétition	0.353	0.359
Synonymie	0.349	0.358
Voisinage	**0.329**	**0.336**

TABLE 1 – Performances de la segmentation thématique selon les liens pris en compte

Ces résultats sont corrects compte-tenu de la difficulté de la tâche : chacune des approches permet une segmentation significativement meilleure que le hasard. Globalement, les résultats

6. Pour chaque texte, une variation de ± 3 ruptures par rapport à la référence est autorisée.

observés avec les différents types de liens cohésifs sont assez proches. L'utilisation des liens de voisinage semble justifiée, puisqu'elle apporte les meilleures performances dans cette expérience, alors que les synonymes ne se démarquent que très peu de l'approche basique par répétitions. Il est donc confirmé que les voisins permettent une détection plus fine de la cohésion lexicale, du moins selon l'étalon que nous avions choisi : la performance d'un système de segmentation thématique.

5 Conclusions et perspectives

L'objectif de cette étude était de montrer la pertinence du voisinage distributionnel pour la détection de la cohésion lexicale. Nous avons à cette fin impliqué les voisins recensés par notre ressource dans un système de segmentation thématique. Les résultats obtenus montrent un apport significatif de la ressource mobilisée. Ainsi, nous avons vu qu'une ressource obtenue grâce à l'analyse distributionnelle présente des avantages que n'ont pas les ressources traditionnelles. Cette expérience mériterait d'être approfondie ; nous aimerions notamment comparer les voisins avec une ressource plus similaire, comme par exemple avec des collocations ; il serait également intéressant d'étudier les possibilités de combinaisons de ressources.

Comme nous l'avons indiqué en introduction, la segmentation thématique n'est pas pour nous une fin en soi. Si nous nous sommes ici donné pour but de repérer des zones cohésives, c'est avant tout pour confronter différentes méthodes de détection des liens lexicaux. En effet, le projet VOILADIS s'inscrit dans une démarche résolument axée vers une analyse du discours qui va au-delà du simple découpage thématique. En nous appuyant sur une ressource obtenue grâce à l'analyse distributionnelle, nous espérons mettre au point une méthode de détection des liens de cohésion lexicale assez efficace pour nous permettre de capter des particularités dans les fonctionnements discursifs des textes, à l'instar des *topic opening* et *topic closing* que repère (Hoey, 1991) en se basant sur la partie du texte vers laquelle pointent les liens cohésifs.

Références

BEEFERMAN D., BERGER A. & LAFFERTY J. (1997). Text segmentation using exponential models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, p. 35–46, Providence.

BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical models for text segmentation. *Mach. Learn.*, **34**(1-3), 177–210.

BOLSHAKOV I. A. & GELBUKH A. (2001). Text segmentation into paragraphs based on local text cohesion. In *TSD '01 : Proceedings of the 4th International Conference on Text, Speech and Dialogue*, p. 158–166, Zelezna Ruda.

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9^e^conférence sur le Traitement Automatique de la Langue Naturelle*, Nancy.

BOURIGAULT D. (2007). *Un analyseur syntaxique opérationnel :* SYNTEX. Habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

BOURIGAULT D. & GALY E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. In *4^es^Journées de la linguistique de corpus*, p. 163–174, Lorient.

BRUNN M., CHALI Y. & PINCHAK C. J. (2001). Text summarization using lexical chains. In *Proceedings of the Document Understanding Conference (DUC 2001)*, p. 135–140, Nouvelle Orléans.

CALLAN J. P., CROFT W. B. & HARDING S. M. (1992). The inquery retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, p. 78–83.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, p. 26–33, San Francisco.

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.

FABRE C. & BOURIGAULT D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Actes de la 13econférence sur le Traitement Automatique de la Langue Naturelle*, Louvain.

FERRET O. (2002). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In *Actes de TALN 2002*, p. 155–165, Nancy.

HALLIDAY M. A. K. & HASAN R. (1976). *Cohesion in English*. Longman (Londres).

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

HERNANDEZ N. (2004). *Description et détection automatique de structures de textes*. PhD thesis, Université Paris-Sud.

HOEY M. (1991). *Patterns of lexis in text*. Oxford University Press (Oxford).

KOZIMA H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, p. 286–288, Columbus.

LIN D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, p. 296–304, Madison.

LIN M., NUNAMAKER JR. J. F., CHAU M. & CHEN H. (2004). Segmentation of lecture videos based on text : a method combining multiple linguistic features. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Hawaii.

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.

MORRIS J. & HIRST G. (2004). Non-classical lexical semantic relations. In *Proceedings of the HLT Workshop on Computational Lexical Semantics*, p. 46–51, Boston.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**, 1–19.

PÉRY-WOODLEY & SCOTT, Eds. (2006). *Discours et Document : traitements automatiques. Numéro thématique*, volume TAL 47(2).