Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes

[validation of a methodology for the study of segmentation markers in large corpora]

Piérard Sophie et Bestgen Yves

Université catholique de Louvain
Place Cardinal Mercier, 10
B-1348 Louvain-la-Neuve

sophie.pierard@psp.ucl.ac.be

RÉSUMÉ. *Cette recherche a pour objectif la validation d'une méthodologie pour l'étude de marqueurs de la segmentation dans un grand corpus de textes. Deux indices signalant plus ou moins efficacement les ruptures thématiques dans un texte sont proposés. Le premier s'appuie sur la présence de marques de paragraphe et emploie le rapport des chances pour identifier les meilleurs marqueurs. Le second prend en compte la cohésion lexicale par l'entremise d'un indice issu de l'analyse sémantique latente. Ces deux indices ont été appliqués principalement à l'étude des expressions adverbiales temporelles dans des textes littéraires. Les analyses effectuées confirment une série d'hypothèses linguistiques au sujet de la fonction de marqueur de la segmentation de ces expressions.*

ABSTRACT. *This research aims at validating a methodology for the study of segmentation markers in large corpora. Two indices signalling a thematic break in a text are proposed. The first is based on the presence of a paragraph mark and employs the odds ratio to identify the best markers. The second takes into account lexical cohesion between sentences via an index resulting from latent semantic analysis. These two indices were applied mainly to the study of temporal adverbial expressions in literary texts. The analyses carried out confirm a series of linguistic hypotheses about the segmentation function of temporal adverbials.*

MOTS-CLÉS : *marqueurs de segmentation, adverbiaux temporels, paragraphe, analyse sémantique latente, expressions référentielles.*

KEYWORDS : *segmentation markers, temporal adverbials, paragraph, latent semantic analysis, referential expressions.*

TAL. Volume 47 – n° 2/2006, pages 89 à 110

1. Introduction

Cette recherche s'inscrit dans la ligne des travaux d'exploitation d'indices de surface et de marqueurs afin de mettre en évidence la structure d'un texte. Un texte n'est pas qu'une suite linéaire de phrases. Il se divise en segments regroupant des ensembles de phrases localement cohérentes. Les relations entre les segments renvoient, quant à elles, à la cohérence globale. La structure du discours est déterminée par ces deux types de cohérence. Identifier cette structure est essentiel pour garantir le succès de la communication entre un auteur et un lecteur. Pour établir la cohérence locale, le lecteur peut s'appuyer sur les marques référentielles (anaphores) et les marques relationnelles (connecteurs) que l'auteur, en vertu du principe de coopération de Grice (1975), est supposé insérer dans son texte. À elles seules, ces marques sont néanmoins insuffisantes, car elles n'apportent des informations qu'à propos des connexions positives dans un texte. Elles sont de peu d'utilité lorsqu'il y a une rupture thématique dans le texte correspondant à la transition entre deux segments (digression, un nouvel épisode,...). Dans cette situation, des marqueurs qui indiquent les ruptures au niveau de la cohérence sont nécessaires pour aider le lecteur/auditeur à ajuster sa représentation mentale du texte. Ce sont précisément les éléments linguistiques qui signalent ces ruptures que nous étudions. Une meilleure connaissance de ceux-ci permettra d'améliorer l'intelligibilité de textes puisque l'usage adéquat de ces marqueurs souligne la structure d'un texte (Givón, 1995 ; Halliday et Hasan, 1976 ; van Dijk et Kintsch, 1983). Ces mêmes marques sont des sources d'information potentiellement très intéressantes pour les systèmes de segmentation automatique des textes comme l'ont montré Passonneau et Litman (1997) et Beeferman, Berger et Lafferty (1999).

Dans le cadre qui vient d'être tracé, cette recherche vise les objectifs suivants. Tout d'abord, nous mettons à l'épreuve et validons une méthodologie pour l'étude de ces marqueurs de la segmentation dans de grands corpus de textes, basée sur l'identification automatique des ruptures thématiques et sur l'analyse de celles-ci en fonction de la présence de candidats-marqueurs. Si les éléments linguistiques qui signalent la structure du discours retiennent depuis de nombreuses années l'attention des chercheurs, les arguments empiriques présentés pour les soutenir sont généralement issus de l'analyse d'un nombre très réduit de textes. La raison en est que l'étude de ces marqueurs présuppose la connaissance de la structure. Classiquement, celle-ci est obtenue par une analyse linguistique fine (pour un exemple récent, voir Asher, Denis et Reese, 2006) ou par le recours à des juges auxquels on demande d'indiquer les ruptures thématiques qu'ils perçoivent (pour un exemple récent, voir Bestgen et Piérard, 2006). La complexité et le coût de ces procédures manuelles rendent l'étude de grands corpus impraticable. Or, nombre de marqueurs linguistiques de la structure, considérés individuellement, ont un taux d'occurrences très faible, imposant justement l'analyse de grands corpus. Pour dépasser cette limitation, il est nécessaire de s'appuyer sur des techniques qui permettent de localiser automatiquement les occurrences de candidats-marqueurs, mais aussi de déterminer tout aussi automatiquement la structure thématique des

textes. Si la première exigence peut être rencontrée par des techniques bien établies en traitement automatique du langage comme l'extraction d'expressions régulières, la mise en évidence de la structure est une problématique encore loin d'être résolue comme le montre, par exemple, les travaux réalisés dans le cadre de DEFT06 dont Hurault-Plantet, Jardino et Berthelin, (2006) et Widlöcher *et al.*, (2006). Nous proposons deux indices susceptibles de combler partiellement cette lacune. Ceux-ci permettent en effet de déterminer si un élément linguistique a tendance à apparaître plus souvent en situation de continuité ou de discontinuité thématique. Le premier indice, très simple puisqu'il s'appuie sur la marque de paragraphe, connaît actuellement un regain d'intérêt (Filippova et Strube, 2006 ; Hœy, 2005 ; Piérard et Bestgen, 2005). Le second indice est issu de développements récents en segmentation automatique de textes (Choi, Wiemer-Hastings et Moore, 2001).

Afin de jauger l'efficacité de ces indices, nous les avons employés pour étudier la fonction de marqueurs de la segmentation des expressions adverbiales temporelles. Il s'agit, et c'est le deuxième objectif de la recherche, de tester plusieurs hypothèses issues de travaux linguistiques à propos des conditions qui déterminent l'efficacité de ces expressions en tant que marqueurs de la segmentation (Charolles, Le Draoulec, Pery-Woodley et Sarda, 2005 ; Le Draoulec et Pery-Woodley, 2005 ; Vonk, Hustinx et Simons, 1992). Tout particulièrement, nous montrons que toutes les expressions temporelles insérées en tête de phrase n'ont pas la même efficacité comme marqueurs de la structure et que ces mêmes expressions voient leur efficacité s'accroître lorsqu'elles sont employées simultanément avec certaines expressions référentielles. Ces résultats permettent de confirmer l'intérêt des recherches de détection automatique des ruptures thématiques basées sur le cumul d'indices.

Dans la suite, nous présentons les deux types de marqueurs de la structure étudiés : les adverbiaux temporels et les expressions référentielles. Nous détaillons les deux indices qui permettent de déterminer si ces expressions s'observent plus fréquemment ou non en situation de ruptures thématiques. Ensuite, nous présentons le corpus analysé et les résultats obtenus. Ces résultats commencent par l'observation d'un fait assez large (les adverbiaux temporels) pour aboutir à une analyse d'un fait plus précis, à savoir l'emploi simultané d'expressions temporelles très efficaces et d'expressions référentielles.

2. Les adverbiaux temporels comme marqueurs de la structure

Tant en linguistique qu'en psycholinguistique, une série de recherches ont montré que les individus qui produisent un discours (à l'oral ou à l'écrit) utilisent la ponctuation, des expressions référentielles plus spécifiques que nécessaire et des expressions adverbiales pour en marquer la segmentation (Chafe, 1979 ; Costermans et Bestgen, 1991 ; Grimes, 1975 ; Longacre, 1979 ; Marcu, 2000 ; Schiffrin, 1987 ; Virtanen, 1992 ; Vonk *et al.*, 1992).

La marque de paragraphe (l'alinéa) est le prototype des marqueurs de segmentation à l'écrit. Bien qu'il remplisse d'autres fonctions discursives (Stark, 1988), il est le moyen le plus courant pour signaler un changement de thème dans un texte (Fayol et Abdi, 1988). La fonction de segmentation de certaines expressions référentielles a elle aussi été démontrée. Les auteurs font référence à différentes entités dans leur texte ou leur discours au moyen de nombreuses expressions différentes allant de l'ellipse (pas d'anaphore) à la reprise du nom, en passant par le pronom (Givón, 1983). Plusieurs facteurs entrent en jeu lors du choix d'une expression référentielle ; parmi ceux-ci, la présence d'un changement d'épisode est très importante. Les auteurs utilisent des expressions plus explicites que nécessaire lors d'un changement de thème (Fox, 1987 ; Hofmann, 1989 ; Vonk *et al*., 1992).

De tous les dispositifs aptes à signaler les ruptures thématiques, c'est très probablement les expressions adverbiales qui ont reçu le plus l'attention. Pour certains auteurs (Brown et Yule, 1983, pp. 95-100 ; Chafe, 1984 ; Longacre, 1979, p. 117-118 ; van Dijk, 1982, p. 181), les adverbiaux temporels et spatiaux, introduits en tête de phrases, sont des « indices grammaticaux » qui mettent en évidence le début d'un nouvel épisode dans une narration. Par l'analyse de narrations, Costermans et Bestgen (1991) et Segal *et al*. (1991) ont observé que l'auteur d'un texte introduit des adverbes temporels comme *Puis* ou *Après* ou des adverbiaux tels que *Vers deux heures* au début des phrases qui introduisent des ruptures dans leurs histoires. D'autres recherches ont souligné l'impact de ces adverbiaux sur les processus mentaux à l'œuvre lors de la compréhension d'un texte. Anderson, Garrod et Sanford (1983), Bestgen et Vonk (1995) ou encore Zwaan (1996) ont montré que des adverbiaux comme *Une heure plus tard* signalent un changement dans la narration et conduisent le lecteur à initialiser la construction d'une nouvelle section dans sa représentation mentale du discours.

Ces travaux mettent l'accent sur le rôle de marqueur de segmentation des adverbiaux temporels. Il importe toutefois de noter que ceux-ci remplissent des fonctions de mise en évidence de la structure d'un texte bien plus complexes que le simple signalement d'un changement de thème. Ceci est particulièrement explicite dans les travaux de Charolles (1997) et Charolles *et al*. (2005) à propos des expressions qui introduisent les cadres du discours. Il s'agit d'adverbiaux extra-prédicatifs antéposés qui indexent les informations qu'ils préfixent en fonction d'un critère qui peut être temporel (*En 2004*), mais aussi spatial (*En France*), ou encore énonciatif (*Selon le secrétaire général*). La portée de ces adverbiaux peut se limiter à une phrase, mais aussi s'étendre au-delà et donc gouverner l'interprétation d'un segment de textes. Ces cadres contribuent à subdiviser et à répartir les informations apportées par le discours au fur et à mesure de son développement. Ils participent de la sorte à son organisation. La fonction discursive des adverbiaux ne se limite pas à signaler une rupture thématique, mais favorise aussi la bonne interprétation du texte en gérant les opérations de mobilisation des connaissances requises pour l'interprétation des relations entre propositions (Charolles, 1997).

Dans cette recherche, seule l'étude de la fonction de marqueur de segmentation des expressions adverbiales temporelles est abordée. Plusieurs arguments justifient cette décision. Tout d'abord, la fonction de marqueur de segmentation est première en ce sens qu'un adverbial ne peut gouverner l'interprétation d'un segment de texte que s'il établit d'abord la présence d'une rupture. Ensuite, Le Draoulec et Péry-Woodley (2005) ont observé que, dans un corpus narratif, les adverbiaux temporels fonctionnaient rarement comme des indicateurs de cadre à proprement parler parce qu'il était souvent difficile de déterminer la fin du cadre en question (sa frontière droite). En d'autres mots, on trouve rarement dans ce type de textes des chaînes d'adverbiaux (Virtanen, 1992) qui structurent le texte en fonction de cette seule dimension temporelle. Enfin, et plus pragmatiquement, l'étude de la fonction cadrative des expressions adverbiales requiert une analyse linguistique fine du texte, actuellement difficile à automatiser, alors que celle de la fonction de marqueur de la segmentation est plus aisément automatisables.

3. Les expressions référentielles comme marqueurs de la structure

En plus de répondre à la question de la fonction de marqueur du discours des adverbiaux temporels, cette étude vise un objectif plus spécifique : étudier les relations entre deux types de marqueurs de la segmentation d'un texte que sont les adverbiaux temporels et les expressions référentielles (nom propre, pronom, nom avec déterminant indéfini, défini,...). Des expressions telles que le pronom personnel sont utilisées dans des situations de continuité de thème. En revanche, des expressions nominales (comme « Jacky » mais aussi « le pharmacien »), lorsqu'elles sont utilisées alors que l'accessibilité à l'antécédent est forte, indiquent une transition vers une nouvelle unité du discours. Les expressions nominales sont donc des signaux de changement de thème lorsqu'elles sont employées alors que le contexte ne le nécessite pas (Asher *et al.*, 2006 ; Vonk *et al*., 1992). Vonk *et al*. (1992) ont demandé à leurs participants d'écrire une suite à de courtes histoires de deux lignes mettant en scène un personnage. Dans l'une des conditions expérimentales, les chercheurs imposaient aux participants d'écrire une suite en rupture ou en continuité thématique par rapport au début du texte. Ils ont observé que les ruptures de thème étaient liées à l'emploi d'anaphores plus spécifiques que nécessaire, c'est-à-dire d'anaphores nominales. De plus, ils ont observé que lorsqu'il y a un changement de thème dans une narration, l'auteur a tendance à employer soit une expression temporelle en début de phrase et un pronom, soit un nom seul. Ils expliquent cette observation en soutenant que la présence d'un marqueur temporel de la segmentation réduit les chances d'observer une expression référentielle plus spécifique que nécessaire. Il n'y aurait donc pas d'emploi simultané de ces deux dispositifs qui indiquent un changement de thème. Ces résultats ont été obtenus au travers d'une tâche relativement artificielle (imposer aux participants de produire des suites en continuité ou en rupture thématique). Un des objectifs de notre recherche est de déterminer si l'emploi simultané ou non de ces

deux types de marqueurs peut être mis en évidence par l'analyse d'un grand corpus de textes.

4. Indices de continuité/discontinuité thématique

Afin de pouvoir analyser le fonctionnement des marqueurs de la segmentation dans de grands corpus de textes, nous proposons deux indices qui permettent de déterminer si une expression linguistique a tendance à apparaître plus souvent ou non en situation de discontinuité thématique. Le premier indice s'appuie sur la présence d'un changement de paragraphe ; le second indice est issu de développements récents en segmentation automatique de textes.

4.1. *Le changement de paragraphe*

En premier lieu, un indice qui traduit, au moins partiellement, les intentions de l'auteur d'un texte a été utilisé : les changements de paragraphe (ou alinéas). L'auteur d'un texte est en effet censé les introduire pour signaler une discontinuité thématique (Hœy, 2005 ; Hofmann, 1989 ; Longacre, 1979).

Nous proposons donc de déterminer si les candidats marqueurs de ruptures thématiques sont plus fréquents dans les phrases qui commencent un nouveau paragraphe plutôt que dans les phrases qui n'en commencent pas. Pour comparer statistiquement ces données, nous avons choisi d'employer l'indice du rapport des chances (RC) qui est associé au classique test du Chi² (Howell, 1998 ; Piérard et Bestgen, 2005 ; Rogati et Yang, 2002). Comparé au Chi², le rapport des chances présente l'avantage de permettre les comparaisons entre expressions parce qu'il n'est pas affecté par leurs fréquences inégales (Howell, 1998, p. 182). Il s'agit, par exemple, du rapport entre la chance qu'une phrase contenant une expression temporelle arrive en tête de paragraphe par rapport à celle qu'une phrase ne contenant pas d'expression temporelle arrive en tête de paragraphe. Ce rapport des chances est calculé sur la base d'une table de contingence comme présentée dans le tableau 1.

Un rapport des chances inférieur à 1 indique donc qu'un candidat marqueur apparaît plus souvent en milieu de paragraphe qu'en début. C'est le cas des expressions indiquant une continuité thématique (comme par exemple, le déterminant possessif, Piérard et Bestgen, 2005). Un rapport des chances supérieur à 1 indique l'inverse : le candidat marqueur apparaît plus souvent en début de paragraphe qu'en milieu ; il signale alors une discontinuité de thème. Plus le rapport des chances est différent de 1 et plus cette différence dans la distribution de l'expression en fonction de sa position est importante. Afin de déterminer à partir de quelle valeur un rapport des chances est suffisamment éloigné de 1 pour pouvoir

être considéré comme statistiquement significatif, on emploie le test du Chi² dont la formule est la suivante :

$$\text{Chi}^2 = \sum \frac{(observé - attendu)^2}{(attendu)^2}$$

Le tableau 1 présente un exemple de RC, basé sur des données recueillies lors de cette étude. Il indique qu'une phrase qui contient une expression temporelle a 1,84 fois plus de chance d'être en tête de paragraphe qu'ailleurs dans le paragraphe, le Chi² étant significatif (Chi²(1)=268,5, p<0,0001).

	Tête de paragraphe	Non tête de paragraphe	Total
Temporel	1 194	1 871	3 065
Non temporel	26 740	77 063	103 803
Total	27 934	78 934	106 868

Tableau 1. *Table de contingence pour les phrases contenant ou non des expressions temporelles.*

$$RC = \frac{1194 / 1871}{26740 / 77063} = 1{,}84$$

L'inconvénient majeur de ce premier indice est que les paragraphes remplissent d'autres fonctions discursives comme la mise en évidence d'un élément du texte (Brown et Yule, 1983 ; Stark, 1988). Il est donc utile de le corroborer par un second indice basé sur la cohésion lexicale entre les phrases d'un texte, indice qui a, entre autres, été employé avec succès pour segmenter automatiquement des textes (Bestgen, 2006 ; Choi *et al.*, 2001).

4.2. *Indice de cohésion lexicale – Analyse Sémantique Latente*

Ce second indice est issu de l'analyse sémantique latente (ASL), une technique mathématique qui vise à extraire un espace sémantique de très grande dimension à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes (Deerwester, Dumais, Furnas, Landauer et Harshman, 1990 ; Landauer et Dumais, 1997). Comme le souligne Landauer, Foltz et Laham (1998), cette technique peut être vue de deux manières. À un niveau théorique, elle peut servir de base pour développer des simulations des processus psycholinguistiques à l'œuvre

lors de la compréhension du langage, incluant, par exemple, un « modèle computationnel » du traitement des métaphores (Kintsch, 2000). À un niveau plus appliqué, c'est une technique permettant d'inférer et de représenter le sens de mots sur la base de leur usage dans des textes, mais aussi d'analyser la cohérence dans des textes (Foltz, Kintsch et Landauer, 1998).

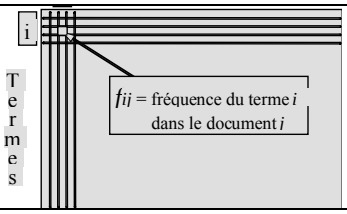
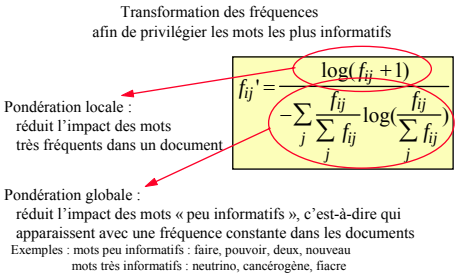
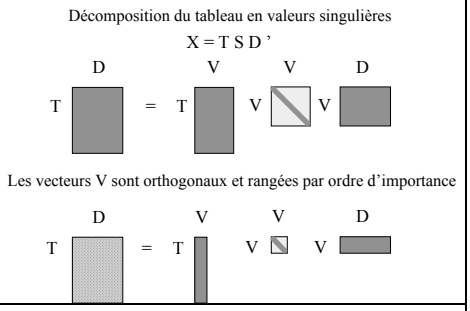
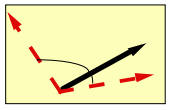
1) Obtention d'un tableau lexical «termes * documents» (nombre d'occurrences de chaque terme dans chaque document)	
2) Transformation des fréquences afin de privilégier les termes les plus informatifs	
3) Décomposition en valeurs singulières Compression de l'information par la sélection des *k* dimensions orthogonales les plus importantes ($100 \leq k \leq 300$) Permet d'obtenir les vecteurs qui représentent les termes dans l'espace comprimé	
4) Emploi : Calculer la proximité sémantique entre des mots ou des segments Le sens d'un mot est représenté par un vecteur La similarité entre deux mots est mesurée par le cosinus entre les vecteurs correspondants (idem pour les segments)	 Soit *X* et *Y* deux vecteurs et θ représentant l'angle entre ces deux vecteur, alors $\cos\theta = \frac{XY}{\|X\| \times \|Y\|}$

Figure 1. *Les étapes d'une analyse sémantique latente.*

La figure 1 résume les différentes étapes qui composent une analyse sémantique latente. Le point de départ de l'analyse est un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chaque segment de textes, un segment pouvant être un paragraphe ou une phrase ou même une suite de mots d'une longueur arbitraire. Ce tableau fait l'objet d'une décomposition en valeurs singulières, une sorte d'analyse factorielle, qui en extrait les dimensions orthogonales les plus importantes. Les milliers de mots caractérisant les documents sont ainsi remplacés par des combinaisons linéaires ou « dimensions sémantiques » sur lesquelles peuvent être situés les mots originaux. Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (Landauer *et al.*, 1998). Dans cet espace, le sens de chaque mot, chaque phrase ou chaque paragraphe est représenté par un vecteur. Pour calculer la similarité sémantique entre deux phrases, on calcule le cosinus entre les vecteurs qui les représentent. Plus deux phrases sont sémantiquement proches, plus leur cosinus est élevé, la valeur maximale étant 1. Un cosinus de 0 indique une absence de similarité. L'intérêt majeur de cet espace sémantique est qu'il permet d'identifier comme similaires deux passages d'un texte même si ceux-ci ont peu de mots en commun.

Nous employons cette technique pour déterminer si une expression linguistique donnée apparaît plus souvent en situation de continuité ou de discontinuité. Plus précisément, nous nous basons sur les cosinus entre la phrase qui inclut cette expression, appelée ici phrase cible (p), et les deux phrases contiguës : celle qui la précède (p-1) et celle qui la suit (p+1). Plus le cosinus entre la phrase cible et celle qui la précède (cos[p-1,p]) est grand, plus la phrase cible est en situation de continuité thématique. Si une phrase cible introduit un changement de thème, c'est le cosinus avec la phrase qui la suit (cos[p, p+1]) qui devrait être plus élevé. L'indice utilisé est donc la différence entre ces deux cosinus (cos[p, p+1] - cos[p-1,p]). Si la différence obtenue est plus grande que 0, elle signifie que la phrase cible est sémantiquement plus similaire à la phrase qui la suit qu'à la phrase qui la précède ; ceci est donc un signe de rupture de thème. En revanche, lorsque la différence obtenue est plus petite que 0, la phrase cible est plus similaire à la phrase qui la précède qu'à la phrase qui la suit et ceci est signe de continuité thématique. L'analyse de variance (ANOVA) est utilisée afin de déterminer si les différences obtenues sont significatives.

Il importe de souligner les limites pratiques de cette technique. Pour calculer la différence de cosinus, il est indispensable de travailler avec des triplets de phrases (la phrase cible, la phrase qui la précède et la phrase qui la suit). Or, les dialogues ont été retirés de nos textes. Certaines phrases cibles ne sont dès lors plus encadrées et leurs différences de cosinus ne peuvent pas être calculées. De plus, lorsque les analyses portent sur des phénomènes linguistiques très spécifiques, l'échantillon de

phrases cibles devient de plus en plus petit, ce qui rend l'analyse de variance moins puissante. Par ailleurs, une série de décisions doivent nécessairement être prises dès la constitution du tableau lexical et tout au long d'une l'analyse sémantique latente, décisions qui peuvent en affecter les résultats. Il s'agit principalement de la sélection des documents (longueurs) et des mots (lemmatisation ou non, suppression des mots les plus rares et les plus fréquents) qui sont pris en compte dans le tableau lexical, de la formule employée pour la pondération des termes et du nombre de dimensions extraites. Il faut donc garder à l'esprit que les résultats d'une ASL sont conditionnés par ces paramètres, ceux que nous avons employés étant présentés dans la section suivante. On notera néanmoins que Bestgen (2004) a montré que l'efficacité d'un algorithme de segmentation automatique de textes, basé sur l'ASL (Choi *et al.*, 2001), n'était que peu influencée par ces paramètres.

5. Collection de textes et prétraitement

La collection est composée de textes littéraires extraits des bases ABU, Intratext et Wordthèque. Elle contient 67 romans du XIX^{e} et XX^{e} siècle (par exemple, « Bouvard et Pécuchet » de Flaubert, « Le Rouge et le Noir » de Stendhal, « Germinie Lacerteux » des frères Goncourt), ce qui totalise approximativement 4 300 000 mots. Les textes ont été découpés en phrases et lemmatisés au moyen du programme *TreeTagger* de Schmid (1994). Les paragraphes qui contenaient des dialogues ont été retirés afin de focaliser les analyses sur l'emploi des indicateurs de la structure à l'écrit. Nous avons également retiré les paragraphes composés d'une seule phrase. En effet, il y a peu de sens de déterminer si une expression temporelle apparaît dans une phrase qui est en tête de paragraphe ou non, lorsque le paragraphe en question est constitué d'une phrase unique. Après ces deux étapes, la collection de textes contient approximativement 107 000 phrases et 2 270 000 mots.

Cette même collection de textes, dans sa version lemmatisée et expurgée des dialogues et des formes fonctionnelles (pronoms, articles,...), a été employée pour construire l'espace sémantique nécessaire pour l'analyse sémantique latente. L'ensemble des textes a été segmenté en unités de 60 à 120 mots composées de phrases entières. Tous les mots dont la fréquence dans la collection de textes était au moins égale à 2 ont été pris en compte. La matrice de cooccurrences ainsi obtenue a été décomposée en valeurs singulières par le programme *Svdpack* (Berry, 1992) et les 300 premiers vecteurs propres ont été conservés.

6. Analyses

Dans un premier temps, nous avons employé une procédure d'extraction d'expressions régulières pour sélectionner de manière automatique les phrases contenant une expression temporelle comme une date (*le 2 avril*), une partie de journée (*dès le matin*), une indication d'heure (*vers midi*), un délai (*une*

heure/semaine/année plus tard), etc. (voir tableau 2). Au total, les phrases sélectionnées représentent 3 % des phrases de la collection de textes.

Catégories	Exemples	Nombre total de cas
vers x heures	à onze heures vers minuit à 18 heures	497
partie de journée	au soir le crépuscule cet après-midi	1613
x suivant	le lendemain le jour suivant l'année suivante	597
saison[1]	en été au printemps	42
tard/après[2]	trois jours plus tard deux jours après un instant après après deux heures	342
date (année)	en 1870	101
date (jour)	le 1er juillet le premier janvier le 18 septembre	193

Tableau 2. *Catégories des marqueurs temporels étudiés.*

Ensuite, nous nous sommes intéressés au positionnement de ces phrases dans les paragraphes. Apparaissent-elles plus souvent en tête de paragraphe ou ailleurs ? Afin de confirmer les résultats obtenus par cet indice, nous en avons employé un autre, issu de l'analyse sémantique latente au moyen duquel nous calculons une différence de proximité sémantique entre les phrases. D'une manière générale, les

[1] En raison du petit nombre de cas de cette catégorie dans le corpus, nous l'avons éliminée dans la suite de nos analyses.
[2] Les expressions de ces deux types étant très similaires, elles ont été regroupées dans la même catégorie.

analyses qui suivent partent de l'observation d'un fait général pour se concentrer sur des phénomènes linguistiques de plus en plus précis.

6.1 *Indice paragraphe*

Nos analyses, illustrées dans la figure 2, indiquent un rapport des chances (RC) de 1,84 ($Chi^2(1)=268,5$, $p<0.0001$) pour les phrases contenant un marqueur temporel. Ce RC indique que les phrases contenant un marqueur temporel ont 1,84 fois plus de chance de se retrouver en tête de paragraphe que celles qui ne contiennent pas de marqueur temporel.

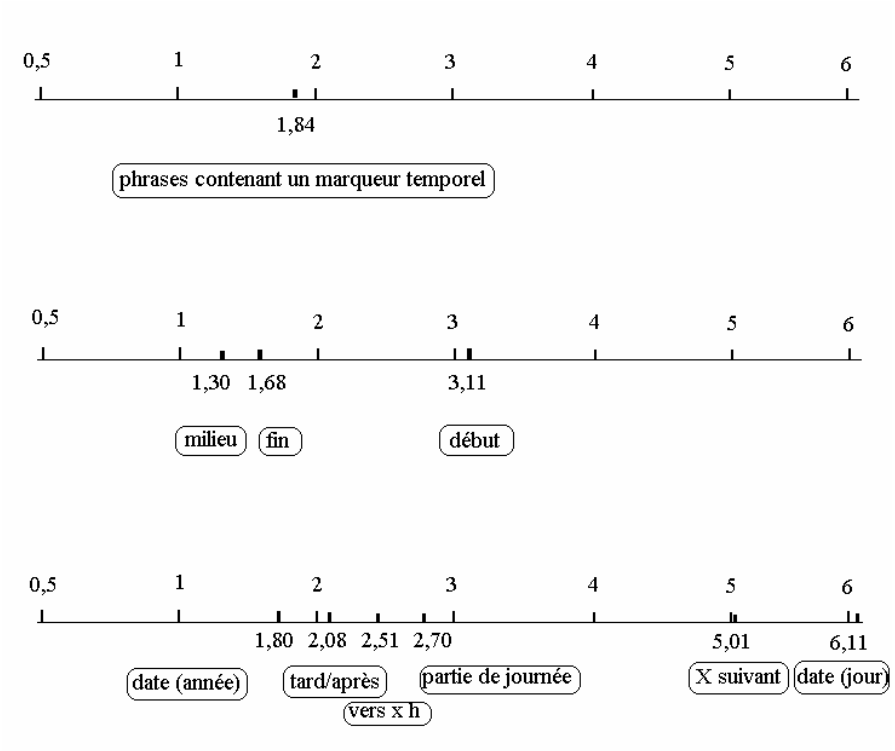


Figure 2. *Rapport des chances des phrases contenant un marqueur temporel, en fonction de la place de celui-ci et selon la catégorie des marqueurs.*

Comme la position initiale des marqueurs temporels semble être la plus efficace pour signaler un changement de thème, nous avons classé ces phrases selon que l'expression temporelle est présente au début, au milieu ou en fin de phrase. Les RC sont respectivement de 3,11, 1,30 et 1,68, tous les Chi² étant très significatifs ($p<0,0001$). Le RC est plus élevé pour les phrases débutant par un marqueur temporel par rapport aux deux autres types de phrases. On note néanmoins que les phrases dans lesquelles l'adverbial n'est pas en tête sont également associées à des changements de paragraphe, même si les RC sont nettement plus faibles par rapport à la position initiale. Ces observations confirment donc l'importance de la position initiale dans la phrase pour qu'une expression temporelle signale le plus efficacement un changement thématique (Charolles, 1997 ; Costermans et Bestgen, 1991 ; Virtanen, 1992).

Selon Costermans et Bestgen (1991), Segal *et al*. (1991), Zwaan (1996) entre autres, toutes les expressions temporelles n'ont pas la même efficacité. Nos analyses confirment cette thèse : certains types de marqueurs apparaissent beaucoup plus souvent que d'autres en tête de paragraphe comme l'indiquent les RC présentés dans la figure 2. Deux types d'expressions obtiennent des RC très élevés. Les phrases débutant par ces expressions sont celles qui, dans notre collection de textes, ont le plus de chance d'apparaître en tête de paragraphe. Ces expressions sont donc les plus efficaces pour indiquer un changement de thème.

6.2 *Différence de cosinus*

Les mêmes analyses que celles décrites ci-dessus ont été effectuées sur la base du second indice, la différence cosinus obtenue grâce à l'analyse sémantique latente. Avant de les rapporter, il est utile de vérifier que ce second indice permet bien de détecter des ruptures thématiques. Dans ce but, nous avons comparé la différence de cosinus pour des phrases qui commencent un paragraphe à la différence de cosinus pour les phrases qui ne commencent pas un paragraphe. Comme attendu, on observe dans la figure 3 que la différence moyenne des cosinus pour les phrases qui commencent un paragraphe est positive (0,0378), ce qui signifie donc qu'elles sont en situation de rupture thématique. En revanche, la différence moyenne des cosinus pour les phrases qui ne commencent pas un paragraphe est inférieure à 0 (- 0,0099), ce qui signifie qu'elles se situent en moyenne plutôt en situation de continuité thématique. L'analyse de variance nous indique que cette différence est très significative ($F(1,83533)=1\,371,14$, $p<0,0001$). En confirmant la fonction de marqueur de rupture du changement de paragraphe, cette analyse indique que cette technique peut être employée pour évaluer la fonction de marqueur de segmentation d'expressions linguistiques.

Si nous distinguons les phrases contenant une expression temporelle des phrases n'en contenant pas, les analyses montrent une différence de cosinus plus élevée pour les phrases qui en contiennent (0,0131) que pour celles qui n'en contiennent pas

(- 0,0002). Ces deux moyennes sont statistiquement différentes (F(1,83533)=18,92, p<0,0001). Les phrases contenant une expression temporelle apparaissent donc plus souvent aux endroits de changements de thèmes dans nos textes que les phrases qui n'en contiennent pas.

Tout comme pour l'indice paragraphe, nous avons classé les phrases selon que l'expression temporelle est présente au début, au milieu ou en fin de phrase. Nous observons que les phrases débutant par un marqueur temporel (0,0268) se distinguent des deux autres types de phrases (respectivement 0,0064 et 0,0057), ces différences sont significatives : F(1,2620)=5,05, p<0,05. Cette analyse confirme l'efficacité de la position initiale de l'expression temporelle comme signal de rupture de thème.

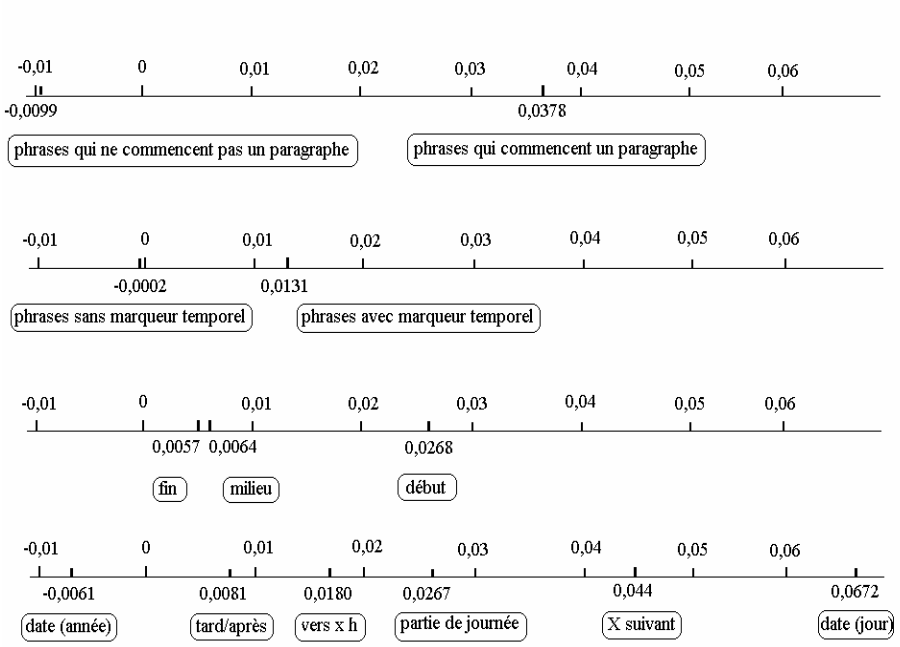


Figure 3. *Différence de cosinus*

Enfin, les expressions temporelles de début de phrases ont été classées en six catégories. De nouveau, les différences sont significatives (F(5,857)=2,45, p<0,05). La comparaison des figures 2 et 3 souligne la très grande similarité entre les résultats obtenus avec les deux indices. Tout particulièrement, les différentes catégories d'expressions temporelles sont ordonnées exactement de la même manière par ces deux indices.

Ces analyses montrent une forte convergence entre les deux indices de la structure. On note néanmoins une exception. Lors de la comparaison de la position des adverbiaux temporels dans les phrases, le RC issu de l'indice paragraphe est plus élevé pour la position finale que pour la position médiane, alors que les différences de cosinus sont très similaires. S'agissant du seul désaccord entre les deux indices et comme nous n'avions pas d'hypothèse a priori à propos de ces positions, il nous semble préférable d'attendre que ce résultat soit reproduit avec d'autres collections de textes avant de tenter de l'expliquer.

L'indice issu de l'ASL semble globalement prometteur même s'il présente plusieurs faiblesses qui seront discutées dans la conclusion. Une de ces limitations réside dans la grande variabilité des cosinus et donc des différences entre les cosinus. Cette variabilité réduit fortement la puissance des analyses de variance, c'est-à-dire leur capacité à mettre en évidence des effets statistiquement significatifs. Comme les analyses qui suivent sont effectuées sur des échantillons de phrases cibles de plus en plus petits, ce manque de puissance nous a conduit à ne plus employer que le seul indice basé sur les paragraphes.

6.3. *Expressions temporelles et expressions référentielles*

Comme déjà annoncé dans nos hypothèses, un des objectifs de cette étude est de démontrer que les expressions temporelles voient leur capacité à signaler une rupture thématique augmenter lorsqu'elles sont associées à des expressions référentielles. Dès lors, nous nous sommes intéressés aux expressions référentielles apparaissant dans les phrases débutant par un marqueur temporel. Le sujet du premier verbe conjugué de chacune de ces phrases a été déterminé au moyen d'une série d'heuristiques syntaxiques. Plus précisément, la procédure traite chaque phrase mot par mot en commençant par le premier. Elle enregistre au fur et à mesure les sujets potentiels grâce à l'étiquetage grammatical effectué par TreeTagger et en prenant en compte la ponctuation et les prépositions pour effacer ou réactiver un sujet potentiel. Lorsque le premier verbe conjugué est rencontré, le sujet le plus probable est sélectionné. Dans un deuxième temps, le sujet est classé dans une des catégories suivantes : Déterminant possessif + SN, Pronom personnel, Déterminant défini + SN, Déterminant indéfini + SN, Déterminant démonstratif + SN, Nom propre. Enfin, une analyse manuelle est appliquée afin d'éliminer les « il » impersonnel (pour une analyse automatique, voir Danlos, 2005). À titre d'exemple, la procédure automatique identifie correctement le sujet du premier verbe conjugué de la phrase suivante comme étant un nom propre : *Cinq minutes après ces navrants aveux, le monde arrivé, les salons pleins, Mme [S Astier S] [V parlait V] et répondait avec une parfaite aisance d'esprit, la mine et la voix heureuses, à me donner la chair de poule.*

Un contrôle manuel effectué sur les 1 028 phrases introduites par une expression temporelle, les seules analysées dans la suite[3], donne un pourcentage d'erreurs inférieur à 5 %, dont la moitié résulte d'un mauvais étiquetage par TreeTagger. Il est à noter que, dans certains cas, la procédure classe la phrase dans la bonne catégorie sans avoir identifié le bon sujet (comme par exemple lorsqu'elle identifie correctement le sujet comme étant un nom propre, mais qu'elle ne sélectionne pas le nom propre adéquat). Ces cas n'ont pas été considérés comme des erreurs puisque notre objectif est d'identifier la catégorie du sujet et non le sujet spécifique. Une limitation de la procédure est qu'elle ne fonctionne pas lorsque le sujet se trouve après le verbe comme dans l'exemple suivant : *Un instant après, sur la place déserte, jonchée de cannes et de chapeaux, [V régnait V] un silence sinistre.* Cette situation est toutefois extrêmement rare dans les phrases analysées.

Nous avons ensuite déterminé si les phrases, débutant par une expression temporelle, dont le sujet est un syntagme avec un article indéfini, un déterminant possessif, etc. étaient plus souvent en tête de paragraphe ou non. Le tableau 3 indique que le nom propre obtient un RC très élevé.

	Nombre de cas	p Valeur du Chi²	RC
Déterminant possessif + SN	15	0,26	0,43
Pronom personnel	310	0,0001	1,89
Déterminant défini + SN	206	0,0001	2,72
Déterminant indéfini + SN	61	0,0001	3,12
Déterminant démonstratif + SN	15	0,0001	4,24
Nom propre	212	0,0001	8,31

Tableau 3. *RC des phrases débutant par un marqueur temporel selon le sujet grammatical (SN signifie syntagme nominal).*

En poussant plus loin l'analyse, on remarque également que le nom propre, présent dans une phrase débutant par un marqueur temporel, est dans 60 % des cas une reprise d'un nom propre cité dans les 10 phrases qui précèdent. Il apparaît que l'utilisation d'un type de marqueurs de rupture comme les adverbiaux temporels n'empêche pas l'utilisation d'autres types de marques comme une expression référentielle plus spécifique tel le nom propre. Ce résultat est en accord avec les observations faites par Hofmann (1989) et Schnedecker (1997). Les indices de segmentation textuelle comme la marque de paragraphe induisent le lecteur à

[3] Parmi celles-ci, 219 phrases ont été placées dans la catégorie « autres sujets » et n'ont pas été prises en compte dans les analyses suivantes parce que, par exemple, le sujet était un pronom personnel de la première personne. Ces phrases ont néanmoins été prises en compte pour évaluer l'efficacité de la procédure d'identification du sujet.

conclure le traitement d'un bloc d'information et à en initialiser un nouveau. Ce nouveau bloc peut débuter par différents types d'expressions et parmi celles-ci, nous pouvons citer les marqueurs temporels. Cette opération implique une accessibilité moins importante des antécédents contenus dans le paragraphe qui vient d'être clôturé. Il est donc nécessaire d'utiliser des marqueurs de plus faible accessibilité comme les noms propres. On peut également constater que le déterminant possessif obtient un RC très faible. Compte tenu de résultats antérieurs (Piérard et Bestgen, 2005) et bien que le Chi² le concernant ne soit pas significatif, mais ceci est lié au petit nombre de cas présents dans l'ensemble des textes, on peut quand même y voir une tendance pour cette expression à être une marque de continuité de thème, réduisant l'effet du marqueur temporel placé en début de phrase.

6.4. *Expressions temporelles les plus efficaces et expressions référentielles*

Nous avons ensuite sélectionné les marqueurs temporels les plus efficaces, à savoir ceux qui obtenaient les meilleurs rapports des chances. Il s'agit des expressions du type « date » (*le 1er juillet*) et les expressions du type *le jour suivant*. Les sujets grammaticaux de ces phrases et leurs RC sont présentés dans le tableau 4. Il apparaît que ces expressions temporelles spécifiques associées à un nom propre obtiennent un RC très important (11,15).

	Nombre de cas	p Valeur du Chi²	RC
Pronom personnel	83	0,0001	3,52
Déterminant défini + SN	57	0,0001	4,18
Déterminant indéfini + SN	18	0,0001	4,44
Nom propre	79	0,0001	11,15

Tableau 4. *RC des phrases débutant par un marqueur temporel du type « date (jour) » ou du type « le jour suivant », selon le sujet grammatical.*

Ce résultat confirme que le cumul de différents types de marques (les expressions temporelles, leurs positions et les expressions référentielles) permet d'obtenir des indications efficaces de rupture thématique.

7. Conclusion

L'objectif principal de cette recherche est de valider une méthodologie pour l'étude de marqueurs de la segmentation dans un grand corpus de textes. Deux indices signalant plus ou moins efficacement les ruptures thématiques dans un texte ont été proposés. Le premier s'appuie sur la présence de marque de paragraphes et

emploie le rapport des chances pour identifier les meilleurs marqueurs. Le second prend en compte la cohésion lexicale par l'entremise d'un indice issu de l'analyse sémantique latente en déterminant si une phrase est sémantiquement moins proche de celle qui la précède que de celle qui la suit. Ces deux indices ont été appliqués principalement à l'étude des expressions adverbiales temporelles dans une collection de textes littéraires.

Les analyses effectuées confirment une série d'hypothèses linguistiques au sujet de la fonction de marqueur de la segmentation de ces expressions. Tant les rapports des chances que l'indice issu de l'ASL mettent en évidence le rôle de signal de rupture des adverbiaux temporels, tout particulièrement lorsque ceux-ci sont insérés en début de phrase. Des analyses complémentaires indiquent que tous les adverbiaux temporels n'ont pas la même efficacité quant au signalement de ruptures dans des textes littéraires. Des expressions comme *le lendemain* y sont beaucoup plus informatives que des expressions comme *en 1980*.

Cette recherche indique aussi que les adverbiaux temporels sont de meilleurs signaux lorsqu'ils couplés avec des expressions référentielles spécifiques. La combinaison d'une expression comme *le 1er juillet* et d'un nom propre comme sujet du premier verbe conjugué de la phrase est dans 80 % des cas associée à un changement de paragraphe.

Un objectif subsidiaire de cette étude est de comparer les deux indices de la structure proposés. Les analyses montrent une forte convergence entre ceux-ci. Non seulement, l'indice basé sur les différences de cosinus, dérivé de l'ASL, est sensible à la présence d'un changement de paragraphes, mais, de plus, il ordonne les types d'expressions temporelles quant à leur efficacité comme marqueur, de la même manière que le fait le rapport des chances. Cet indice semble donc prometteur. Dans sa version actuelle, il est néanmoins moins satisfaisant que l'indice basé sur les paragraphes. Trois explications peuvent être proposées. Tout d'abord, on constate une grande variabilité des cosinus obtenus, rendant les analyses statistiques moins fiables. Une analyse des facteurs responsables de cette variabilité est donc nécessaire et, plus particulièrement, de l'impact de la longueur des phrases sur la taille du cosinus. Ensuite, l'indice de cohésion lexicale a une portée très locale. Il indique seulement qu'une phrase est sémantiquement plus liée avec celle qui la suit qu'avec celle qui la précède ou l'inverse. Le paragraphe est par contre une mesure plus globale puisque l'auteur l'emploie pour segmenter le texte en paquets de phrases qui forment à chaque fois une unité textuelle. Cette faiblesse de l'indice cosinus pourrait être, au moins partiellement, effacée en l'utilisant, non de manière brute, mais au travers d'algorithmes de segmentation automatique comme TextTiling (Hearst, 1997) ou CWM (Choi *et al.*, 2001). Enfin, passer à un nouvel alinéa et introduire un marqueur de la structure sont deux décisions de l'auteur d'un texte. Il n'est donc pas étonnant que ces deux décisions soient fréquemment prises simultanément.

Parmi les principales limitations de cette recherche, il est impossible de ne pas mentionner la question du degré de généralisation des résultats. La collection de textes utilisée est constituée de romans, dans lesquels des repères temporels jalonnent le déroulement de l'histoire afin de situer les actions. Qu'en serait-il dans un autre corpus ? Il serait utile d'effectuer les mêmes analyses dans un corpus d'articles de journaux. Dans ce genre de textes, il n'est pas évident que les expressions temporelles soient associées d'une manière aussi systématique à la présence de changement de thème. Si c'est néanmoins le cas, il serait particulièrement intéressant de déterminer si les différences d'efficacité des différents types d'expressions temporelles se retrouvent dans ce genre de textes ou si la hiérarchie observée dans les œuvres littéraires est spécifique à celles-ci. Une deuxième limitation trouve son origine dans l'unique fonction des expressions temporelles considérées ici, celle de marqueur de la segmentation. Comme indiqué dans l'introduction, ceux-ci remplissent d'autres fonctions, plus complexes, de mise en évidence de la structure d'un texte (Charolles, 1997). L'étude de celles-ci nécessite une analyse linguistique plus fine, qui n'a pas été effectuée dans cette recherche.

Il serait aussi intéressant d'affiner l'analyse des expressions référentielles afin de prendre en compte une catégorisation moins basique que celle que nous avons employée. On peut en effet penser que d'autres facteurs entrent en jeu comme, par exemple, la longueur d'une expression nominale ainsi que l'ont montré récemment Asher et al. (2006) sur la base d'un corpus journalistique. Ces hypothèses rejoignent les analyses de Vonk et al. (1992) qui ont montré que les expressions nominales surspécifiées, et donc les plus longues et les plus détaillées, signalent des ruptures thématiques. Par ailleurs, nos analyses portent sur le sujet du premier verbe conjugué. S'il s'agit fréquemment du sujet de la proposition principale, ce n'est pas toujours le cas. Ce facteur n'a pas été pris en compte.

En conclusion, la méthodologie proposée permet d'envisager l'exploitation du cumul des marqueurs de rupture de thème pour les systèmes de segmentation automatique des textes. Elle autorise plusieurs développements comme l'étude d'autre type de marqueurs (tels les marqueurs spatiaux), ou d'autres combinaisons appliquées à des corpus différents afin de confirmer des hypothèses linguistiques, mais aussi d'effectuer des analyses plus heuristiques.

Remerciements

Yves Bestgen est chercheur qualifié du Fonds national de la recherche scientifique (FNRS). Cette recherche est financée par une « Action de Recherche concertée » du Gouvernement de la Communauté française de Belgique.

8. Bibliographie

Anderson A., Garrod S.C., Sanford A.J., « The accessibility of pronominal antecedents as a function of episode shifts in narrative text », *Quarterly Journal of Experimental Psychology*, 35A, 1983, p. 427-440.

Asher N., Denis P., Reese, B., « Names and pops and discourse structure », *Proceedings of the workshop on constraints in discourse*, Maynooth, 2006, National University of Ireland, p.11-18.

Beeferman D., Berger A., Lafferty J., « Statistical models for text segmentation », *Machine Learning*, 34, 1999, p. 177–210.

Berry M.W., « Large scale singular value computation », *International Journal of Supercomputer Application*, 6, 1992, p. 13-49.

Bestgen Y., « Analyse sémantique latente et segmentation automatique des textes », In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Actes des 7e journées internationales d'analyse statistique des données textuelles (JADT04)* , Louvain-la-Neuve, 10-12 mars 2004, Presses Universitaires de Louvain, p. 171-181.

Bestgen, Y., « Improving text segmentation using Latent Semantic Analysis : A reanalysis of Choi, Wiemer-Hastings and Moore (2001) », *Computational Linguistics*, 32, 2006, p. 5-12.

Bestgen, Y, Piérard S., « Comment évaluer les algorithmes de segmentation automatique? Essai de construction d'un matériel de référence », *Actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN06)*, Leuven, 10-13 avril 2006, Louvain, Presses Universitaires de Louvain, p. 407-414.

Bestgen Y., Vonk W., « The role of temporal segmentation markers in discourse processing », *Discourse Processes*, 19, 1995, p. 385-406.

Brown G., Yule G. , *Discourse analysis*, Cambridge, Cambridge University Press, 1983.

Chafe W., « The flow of thought and the flow of language », In T. Givón (Ed.), *Syntax and Semantics (XII) : Discourse and Syntax*, 1979, New York, Academic Press, p.159-182.

Chafe W., « How people use adverbial clauses », In C. Brugman, M. Macauley (Eds.), *Proceedings of the 10th annual meeting of the Berkeley Linguistics Society*, 1984, Berkeley CA, p.437-449.

Charolles M. « L'encadrement du discours - univers, champs, domaines et espaces », *Cahier de Recherche Linguistique*, 6, 1997, p. 1-73.

Charolles M., Le Draoulec A., Pery-Woodley M-P., Sarda L., « Temporal and spatial dimensions of discourse organisation », *French Language Studies*, 15, 2005, p. 115-130.

Choi F., Wiemer-Hastings P., Moore J., « Latent semantic analysis for text segmentation » , *Proceedings of the conference on empirical methods in natural language processing*, Carnegie Mellon University, 2001, Pittsburgh, PA, p. 109–117.

Costermans J., Bestgen Y., « The role of temporal markers in the segmentation of narrative discourse », *CPC : European Bulletin of Cognitive Psychology*, 11, 1991, p. 349-370.

Danlos L., « ILIMP : outil pour repérer les occurrences du pronom impersonnel il », *Actes de la 12e conférence sur le traitement automatique des langues naturelles (TALN05)*, Dourdan, 6-10 juin 2005, p.123-132.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, 41, 1990, p. 391-407.

Fayol M., Abdi H., « Influence of script structure on punctuation », *Cahiers de Psychologie Cognitive*, 8, 1988, p. 265-279.

Filippova K, Strube M., « Using linguistically motivated features for paragraph boundary identification », *Proceedings of EMNLP 2006*, Sydney, Australia, p.267-274.

Foltz P.W., Kintsch W., Landauer T.K., « The measurement of textual coherence with Latent Semantic Analysis », *Discourses Processes*, 25, 1998, p. 285-307.

Fox B.A., « Morpho-syntactic markedness and discourse structure », *Journal of Pragmatics*, 11, 1987, p. 359-375.

Givón T., « Topic continuity in discourse : quantified cross-language studies », *Typological Studies in Language* #3, Amsterdam, Benjamins, 1983.

Givón T., « Coherence in text vs. coherence in mind », In M.A. Gernsbacher, T. Givón, (Eds), *Coherence in spontaneous text,* Amsterdam, Benjamins, 1995.

Grice H.P., « Logique et conversation », *Communications*, 30, 1975, p. 57-72.

Grimes J.E., *The thread of discourse*. The Hague, Mouton, 1975.

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

Hearst M., « TextTiling : Segmenting text into multi-paragraph subtopic passages », *Computational Linguistics*, 23, 1997, p. 33–64.

Hœy M., *Lexical priming : a new theory of words and language*, London, Routledge, 2005.

Hofmann T.R., « Paragraphs, & anaphora », *Journal of Pragmatics*, 13, 1989, p. 239-250

Howell D.C., *Méthodes statistiques en sciences humaines*, Bruxelles, De Bœck Université, 1998.

Hurault-Plantet M., Jardino M., Berthelin, J.-B., « Ajustement des frontières de segments thématiques détectés automatiquement », *Actes de DEfi fouille de texte (DEFT'06), semaine du document numérique (SDN'06)*, 2006, Fribourg, Suisse.

Kintsch W., « Metaphor comprehension : A computational theory », *Psychonomic Bulletin and Review*, 7, 2000, p. 257-266.

Landauer T.K., Dumais S.T., « A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge », *Psychological Review*, 104, 1997, p. 211-240.

Landauer T.K., Foltz P.W., Laham D., « An introduction to Latent Semantic Analysis », *Discourse Processes*, 25, 1998, p. 259-284.

Le Draoulec, A.,, Péry-Woodley, M.-P., « Encadrement temporel et relations de discours », *Langue Française*, 148, 2005, p. 45-60.

Longacre R. E., « The paragraph as a grammatical unit », In T. Givón (Ed.), *Syntax and Semantics, 12 : Discourse and Syntax,* New York, Academic Press, 1979.

Marcu D., « The rhetorical parsing of unrestricted texts : A surface-based approach », *Computational Linguistics*, 26, 2000, p. 395-448.

Passonneau R.J., Litman D.J., « Discourse segmentation by human and automated means », *Computational Linguistics*, 23, 1997, p. 103-139.

Piérard S., Bestgen Y., « Identification automatique des marqueurs globaux du discours par l'analyse des expressions récurrentes », *Actes de la conférence Phraseology 2005*, Louvain-la-Neuve, 13-15 octobre 2005, p. 343-345.

Rogati M., Yang Y., « High-performing feature selection for text classification », *Proceeding of CIKM'02,* November 4–9 2002, McLean, Virginia.

Schiffrin D., *Discourse markers*, Cambridge, Cambridge University Press, 1987.

Schmid H., « Probabilistic Part-of-speech tagging using decision trees », *Proceedings of international conference on new methods in language processing*, 1994.

Schnedecker C., *Nom propre et chaînes de référence*, Paris, Klincksieck, 1997.

Stark H.A., « What do paragraph markings do? », *Discourse Processes*, 11, 1988, p. 275-303.

van Dijk T., « Episodes as units of discourse analysis », In D. Tannen (Ed.), *Analysing discourse : text and talk*, Washington, Georgetown University Press, 1982.

van Dijk T., Kintsch W., *Strategies of discourse comprehension*, New York, Academic Press, 1983.

Virtanen T., *Discourse functions of adverbial placement in English*, Åbo : Åbo Akademi University Press, 1992.

Vonk W., Hustinx L.G., Simons W.H., « The use of referential expressions in structuring discourse », *Language and Cognitive Processes*, 7, 1992, p. 301-333.

Widlöcher A., Bilhaut F., Hernandez N., Rioult F., Charnois T., Ferrari S., Enjalbert P., « Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte », *Actes de DEfi fouille de texte (DEFT'06), semaine du document numérique (SDN'06)*, 2006, Fribourg, Suisse.

Zwaan R.A., « Processing narrative time shifts », *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 22, 1996, p. 1196-1207.