Partie 2 : les études de corpus

Chapitre 1 : introduction théorique

Les expériences de production de la première partie ne sont pas toujours concluantes. Dès lors, nous avons décidé d'aborder les anaphores possessives autrement et de les observer en contexte naturel, c'est-à-dire en corpus. Les expressions référentielles sont-elles des marqueurs de la structure d'un texte ? Sont-elles employées simultanément avec d'autres marqueurs de la structure ? Où les situer par rapport à d'autres marques de la structure du discours ? Afin de répondre à ces questions, il nous faut d'abord valider une technique qui permet de repérer automatiquement les expressions qui signalent la structure d'un texte : ce sera l'objectif de la première analyse de corpus. Dans l'analyse de corpus 2, nous testerons si cette technique permet de discriminer la capacité de marquage de différents types d'expressions temporelles, comme la littérature l'avance, et nous confronterons cette technique automatique à d'autres indices (automatiques et manuels). Ensuite, dans l'analyse de corpus 3, nous allons comparer la fonction d'indicateur de la structure du discours de différents éléments linguistiques (expressions référentielles et adverbiaux temporels). Pour finir, dans l'analyse de corpus 4, nous allons progressivement mettre en évidence les expressions temporelles les plus efficaces pour signaler une rupture thématique (telle que mesurée par nos indices). Ensuite, nous analyserons les expressions référentielles qui apparaissent au début des phrases contenant ces expressions temporelles afin de mettre en évidence un éventuel usage simultané d'adverbiaux temporels et d'expressions référentielles spécifiques comme le nom propre. En effet, Vonk & al (1992) affirment que des expressions référentielles plus spécifiques que nécessaire signalent un changement de thème. Ils ont également observé, dans leurs expériences, que lorsqu'il y a un changement de thème dans une narration, le scripteur 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

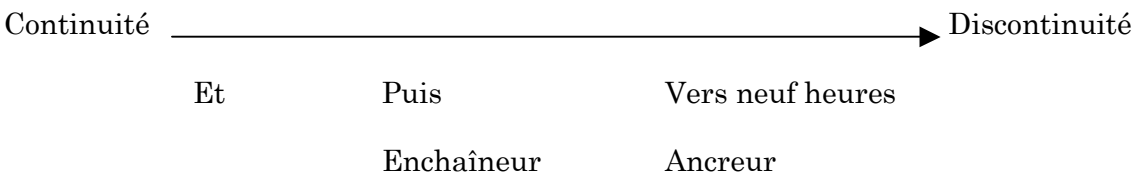
segmentation réduit les chances d'observer une expression référentielle plus spécifique que nécessaire. Il n'y aurait donc pas d'emploi simultané de ces deux dispositifs qui indiquent un changement de thème. Ces résultats ont été obtenus en imposant aux participants de produire des suites en continuité ou en rupture thématique. Nous allons 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.

Dans la suite, nous discutons, sur base de la littérature, la fonction de structuration du discours des expressions temporelles. Ensuite, nous présentons les deux indices que nous avons utilisés ainsi que les corpus. Enfin, les quatre analyses de corpus sont détaillées.

1. Les expressions temporelles comme indice de la structure d'un texte

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 & Bestgen, 1991 ; Fayol & Abdi, 1988 ; Grimes, 1975 ; Longacre, 1979 ; Marcu, 2000 ; Schiffrin, 1987 ; Virtanen, 1992 ; Vonk & al., 1992). De tous ces dispositifs, ce sont très probablement les expressions adverbiales qui ont retenu le plus l'attention. Pour certains auteurs (Brown & Yule, 1983, pp. 95-100 ; Chafe, 1984 ; Longacre, 1979, pp. 117-118 ; van Dijk, 1982, p. 181), les adverbiaux temporels comme "Lundi", "Cet après-midi", "Vers dix heures", introduits en tête de phrase, sont des « indices grammaticaux » qui, dans de nombreuses langues, mettent en évidence les grandes ruptures dans une narration. À l'instar d'Ehrich & Koster (1983), Costermans & Bestgen (1991) ont appelé ce type d'adverbiaux temporels des marqueurs d'ancrage parce qu'ils établissent le début d'une action ou d'un évènement sur une échelle temporelle

absolue extérieure à la narration. Si ces adverbiaux signalent les ruptures les plus importantes dans la narration, d'autres dispositifs comme les adverbes temporels (*puis*, *ensuite)* et le connecteur *et* asymétrique[1] (Lakoff, 1971), sont employés pour mettre en évidence les ruptures moins importantes et même la continuité (Costermans & Bestgen, 1991 ; Fayol, 1986 ; Jisa, 1985 ; Segal, Duchan & Scott, 1991). Cette fonction découle de leur implication dans l'expression des relations temporelles entre des actions ou des évènements successifs dans les narrations. Selon Reichenbach (1947 cité par Le Draoulec & Péry-Woodley, 2003), la manière la plus simple pour un locuteur ou un scripteur d'accomplir cette tâche est d'exprimer deux actions successives en juxtaposant deux phrases : "Je me suis levé. J'ai déjeuné." Cette stratégie indique implicitement que le second évènement suit le premier (Dowty, 1986; Hinrichs, 1986; Partee, 1984). On peut mettre en évidence une continuité entre des actions ou des évènements en introduisant le connecteur *et*. Pour mettre en évidence une discontinuité, un adverbe temporel comme *puis* ou *ensuite* peut être introduit. Costermans & Bestgen (1991), mais aussi Noyau & al. (2005) les ont appelés enchaîneurs (marqueurs séquentiels) parce qu'ils soulignent explicitement qu'un événement se produit après l'événement qui le précède dans la narration. Bestgen & Costermans (1994) ont échelonné quatre types d'expressions allant de la discontinuité thématique la plus importante à la continuité la plus importante : ancreurs > enchaîneurs > aucune marque > et.



Par l'analyse de narrations, Costermans & Bestgen (1991) et Segal & al. (1991) ont observé que l'auteur d'un texte introduit des adverbes temporels comme *Puis*

[1] Le « et » asymétrique impose un ordre de priorité aux phrases qu'il lie ; il est équivalent au « et puis » (Lakoff, 1971 ; 126).

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 & al. (1983), Bestgen & 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 & al. (2005) à propos des expressions qui introduisent les cadres du discours. Il s'agit d'adverbiaux extra-prédicatifs[2] antéposés[3] 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 texte. 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).

[2] « hors phrase »

[3] placés devant

Nous avons étudié ces marques en corpus. Nous avons utilisé à travers nos recherches deux indices : le changement de paragraphe (qui connaît un regain d'intérêt en tant qu'indice de segmentation de textes, voir Filippova & Strube, 2006; Hœy, 2005) et un indice de cohésion lexicale issu de l'Analyse Sémantique Latente (ASL), décrits dans le chapitre suivant.

2. Les indices utilisés dans nos études de corpus

2.1 Le changement de paragraphe

Pour déterminer si une expression a tendance à apparaître plus souvent en situation de continuité ou de discontinuité thématique, nous proposons, en premier lieu, d'employer un indice qui traduit, au moins partiellement, les intentions de l'auteur d'un texte : les changements de paragraphe (ou alinéas). L'auteur d'un texte est en effet censé les introduire pour signaler une discontinuité thématique (Hofmann, 1989 ; Longacre, 1979). Nous proposons donc de déterminer si un élément linguistique apparaît plutôt en début de paragraphes qu'au milieu de ceux-ci en utilisant la technique classique du test du Chi^2 et le rapport des chances qui lui est associé (Howell, 1998 ; Piérard & Bestgen, 2005 ; Rogati & Yang, 2002). Un marqueur de continuité thématique devrait se trouver plus souvent au sein d'un paragraphe alors qu'un marqueur de discontinuité devrait apparaître plus fréquemment en début de paragraphe.

Nous avons choisi d'employer l'indice du rapport des chances (RC) qui est associé au classique test du Chi^2 (Howell, 1998 ; Piérard & Bestgen, 2005 ; Rogati & Yang, 2002). Comparé au Chi^2, 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 23.

Un rapport des chances inférieur à 1 indique 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 & 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 23 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 ($Chi^2(1)=268.5$, $p<0.0001$).

Tableau 23 : Table de contingence pour les phrases contenant ou non des expressions temporelles et formule du rapport des chances associé

	Tête de paragraphe	Non tête de paragraphe	Total
Temporel	1194	1871	3065
Non temporel	26740	77063	103803
Total	27934	78934	106868

$$RC = \frac{1194 / 1871}{26740 / 77063} = 1.84$$

L'inconvénient majeur de ce premier indice est que les paragraphes remplissent d'autres fonctions discursives (Stark, 1988 ; Brown & Yule, 1983), comme par exemple, la mise en évidence d'un élément du texte. Nous proposons donc d'employer parallèlement un indice basé sur la cohésion lexicale qui a, entre autres, été employé avec succès pour segmenter automatiquement des textes (Bestgen, 2005 ; Choi, Wiemer-Hastings & Moore, 2001).

2.2 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 & Harshman, 1990 ; Landauer & Dumais, 1997). Comme le soulignent Landauer, Foltz & 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 (Bestgen, Degand, & Spooren, 2006 ; Foltz, Kintsch & Landauer, 1998).

Comme indiqué dans le paragraphe précédent, l'analyse sémantique latente s'effectue sur de grands corpus de textes. Afin d'illustrer les différentes étapes de la technique, nous allons nous baser principalement sur un petit extrait de trois phrases issu de notre corpus de textes littéraires composés de 67 oeuvres totalisant 4 300 000 mots (voir partie 2, chapitre 4).

« Et plusieurs disaient, dans l' île, que, la nuit, sur les chemins passaient des anges et qu' on mourait pour les avoir vus. Kraken ignorait les amours d' Orberose et de Marcel, car il était un héros, et les héros ne pénètrent jamais les secrets de leurs femmes. Mais, tout en ignorant ces amours, Kraken en goûtait les précieux avantages. »

« L'île aux pingouins », Anatole France

Dans un premier temps, le corpus est transformé en un tableau lexical. Chaque cellule contient la fréquence ($f$ij) avec laquelle chaque terme i (représenté par une ligne) apparaît dans chaque segment j (représenté par une colonne), un segment pouvant être un paragraphe ou une phrase ou même une suite de mots d'une longueur arbitraire.

Pour notre exemple, le segment étudié sera la phrase. Nous allons donc établir un tableau lexical qui va illustrer les cooccurrences de chaque terme dans chacune des trois phrases.

	Phrase 1	Phrase 2	Phrase 3
et	2	2	0
plusieurs	1	0	0
disaient	1	0	0
dans	1	0	0
l	1	0	0
île	1	0	0
que	1	0	0
la	1	0	0
nuit	1	0	0
sur	1	0	0
les	2	3	2
chemins	1	0	0
passaient	1	0	0
des	1	0	0
anges	1	0	0
qu’	1	0	0
on	1	0	0
mourait	1	0	0
pour	1	0	0
avoir	1	0	0
vus	1	0	0
Kraken	0	1	1
ignorait	0	1	1
amours	0	1	1
Orberose	0	1	0
de	0	2	0
Marcel	0	1	0
car	0	1	0
il	0	1	0
était	0	1	0
un	0	1	0
héros	0	2	0
ne	0	1	0
pénètrent	0	1	0
jamais	0	1	0
secrets	0	1	0
leurs	0	1	0
femmes	0	1	0
mais	0	0	1
tout	0	0	1
en	0	0	2
ignorant	0	0	1
ces	0	0	1
goûtait	0	0	1
précieux	0	0	1
avantages	0	0	1

Pré-traitements

Avant de passer à l'analyse à proprement parler, le corpus subit habituellement plusieurs pré-traitements. Si ceux-ci sont facultatifs, ils présentent l'intérêt de réduire le nombre de termes différents et de supprimer des termes souvent peu utiles pour une analyse sémantique latente. Chaque terme portant une marque de flexion (comme un verbe conjugué ou un syntagme nominal au pluriel) est ramené à sa forme de base (ou forme de référence). Dans nos analyses, cette tâche est effectuée au moyen du programme « TreeTagger » de Schmidt (1994). De plus, une série de mots appartenant à des classes fermées (article, pronom, …) ou très fréquents et donc peu informatifs (ils apparaissent souvent et dans tous les textes ; par exemple : avoir, tout) sont supprimés. Après ces pré-traitements linguistiques, voici à quoi ressemble le tableau lexical pour notre exemple.

	Phrase 1	Phrase 2	Phrase 3
dire	1	0	0
île	1	0	0
nuit	1	0	0
chemin	1	0	0
passer	1	0	0
ange	1	0	0
mourir	1	0	0
voir	1	0	0
Kraken	0	1	1
ignorer	0	1	1
amour	0	1	1
Orberose	0	1	0
Marcel	0	1	0
héros	0	2	0
pénétrer	0	1	0
secret	0	1	0
femme	0	1	0
goûter	0	0	1
précieux	0	0	1
avantage	0	0	1

Pondération des fréquences brutes

Ensuite, une transformation est appliquée aux fréquences contenues dans le tableau lexical afin de privilégier les termes les plus informatifs, c'est-à-dire les termes qui sont le plus liés à des segments spécifiques. L'idée sous-jacente est qu'un terme qui apparaît souvent dans quelques segments et presque jamais dans tous les autres segments, est plus "informatif" qu'un terme qui apparaît avec une fréquence relativement constante dans presque tous les segments. Par informatif, on veut dire 1) que ce terme représente une partie plus importante du contenu du segment s'il y est représenté plusieurs fois et 2) que sachant que ce terme est présent dans un segment, il est plus facile d'identifier le segment en question. La transformation recommandée par Landauer et al. (1998) est donnée ci-dessous. Dans cette formule, l'indice *i* correspond aux termes et l'indice *j* au segment. *fij* correspond à la fréquence du terme *i* dans le segment *j*. *Pij* correspond à la fréquence du terme *i* dans le segment *j* divisé par la fréquence totale du terme dans le corpus.

Le numérateur permet de pondérer le terme en fonction de son importance dans le segment particulier dans lequel il apparaît. Un terme qui apparaît plusieurs fois dans un segment est plus informatif qu'un terme qui n'apparaît qu'une seule fois. Il est donc pertinent de prendre en compte la fréquence du terme dans le segment. Toutefois, il est peu probable qu'un terme qui apparaît 2 fois dans un segment soit 2 fois plus informatif qu'un terme qui n'apparaît qu'une fois, et qu'un terme qui apparaît trois fois soit trois fois plus informatif. Aussi prend-on la fréquence du terme dans le segment, mais après avoir appliqué une transformation logarithmique à cette fréquence de façon à limiter l'impact de celle-ci. Comme on a f=1 -> log10(1+1)=0.30 ; f=2 -> log10(2+1)= 0.48 ; f=3 -> log10(3+1)= 0.60, après transformation logarithmique, une fréquence de 3 n'est que deux fois plus importante qu'une fréquence de 1.

Le dénominateur pondère la fréquence du terme en fonction de l'information qu'il apporte dans le corpus en général. À chaque terme correspond une seule valeur qui est donc appliquée à chaque fréquence correspondant à ce terme. L'idée sous-jacente est la suivante. Si un terme apparaît dans tous ou presque tous les

segments, ce terme sera pour nous peu informatif puisqu'il cooccurre avec tous ou presque tous les termes du corpus, et donc son occurrence ne varie pas en fonction du thème abordé dans les différents segments. Par contre, un terme qui n'apparaît que dans quelques segments aborde beaucoup plus d'informations puisqu'il ne cooccurre qu'avec un petit nombre d'autres termes. La pondération employée égale à l'inverse de l'entropie telle que définie par Shannon[4] dans la théorie de l'information. Elle nous indique dans quelle mesure le fait de savoir qu'un terme est présent dans un segment nous permet de déterminer de quel segment il s'agit. Deux brefs exemples permettront de comprendre l'impact de cette pondération. Considérons un mini-corpus de 10 segments et un terme qui apparaît au total 10 fois. Si ce terme est présent dans chaque segment, sa pondération sera $-[1/10 \times \log10^{5}(1/10)] \times 10 = 1$. On divisera donc l'ensemble des f_{ij} correspondant à ce terme par 1. Si, par contre le terme n'apparaît que dans deux segments avec une fréquence de 5, sa pondération sera $-[0.5 \times \log10(0.5)] \times 2 = 0.30$. On divisera donc l'ensemble des f_{ij} correspondant à ce terme par 0.3, ce qui revient à les multiplier par 3.32. Un terme qui n'apparaît que dans un seul segment reçoit une pondération de 0 et n'est donc pas pris en compte dans les analyses en raison de la division par 0 qu'il impose. On voit que la fréquence joue un rôle important dans cette pondération. Un terme très fréquent apparaît souvent dans plus de segments qu'un terme rare. Dans notre corpus littéraire, les termes comme « aimer », « femme », ou encore « homme » sont peu intéressants parce qu'ils apparaissent très souvent dans de nombreux segments. Rien de plus normal de parler d'amour, d'homme et de femme dans des romans…Par contre, les termes « abricot », « slave » et « abbesse » sont très intéressants pour nous, car ils apparaissent beaucoup moins fréquemment.

---

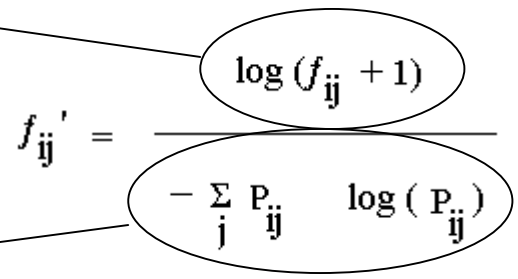
[4] L'entropie de Shannon est une fonction mathématique qui correspond à la quantité d'information contenue ou délivrée par une source d'information. Elle peut être vue comme mesurant la quantité d'*incertitude* liée à un évènement aléatoire, ou plus précisément à sa distribution.

[5] On aurait pu employer les logarithmes en base 2 comme dans la théorie de l'information, mais cela ne changerait rien à la démonstration.

L'amélioration apportée par ce genre de transformation a été démontrée en recherche automatique d'information (Landauer & al., 1998)).

pondération locale :

réduit l'impact des termes très fréquents

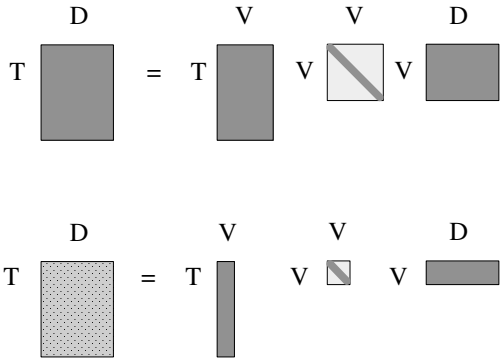


pondération globale :

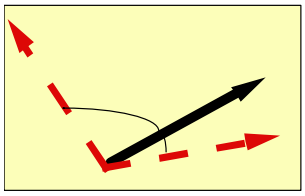
réduit l'impact des termes peu informatifs

Décomposition en valeurs singulières

Enfin, 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. Cette procédure mathématique, ancienne et à la base de nombreuses procédures d'analyses multidimensionnelles de données. La matrice de base est décomposée en trois matrices de telle manière que le produit de celles-ci permet de recomposer la matrice originale. L'une d'entre elles décrit les lignes originales comme étant des vecteurs, une autre décrit les colonnes de la même manière et enfin la troisième est une matrice diagonale contenant les valeurs échelonnées de manière à ce que lorsque les trois matrices sont multipliées, la matrice originale soit reconstituée (Landauer & al., 1998). Si on élimine les valeurs les plus faibles de cette dernière matrice, le produit des trois matrices donnera, non la matrice originale, mais la meilleure approximation possible de celle-ci (au sens des moindres carrés). Cette opération est équivalente à celle qui consiste à ne conserver que les "facteurs" les plus importants d'une analyse en composantes principales.



Les milliers de termes caractérisant les documents sont ainsi remplacés par des combinaisons linéaires, ou dimensions sémantiques, sur lesquelles peuvent être situés les termes 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 & al., 1998).



Soit X et Y deux vecteurs et θ représentant l'angle entre ces deux vecteur, alors

$$\cos\theta = \frac{XY}{\|X\| \times \|Y\|}$$

Dans cet espace, le sens de chaque terme, de chaque phrase ou de chaque paragraphe est représenté par un vecteur. Pour calculer la similarité sémantique entre deux phrases, on calcule habituellement 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.

Pour reprendre notre exemple :

Kraken ignorait les amours d' Orberose et de Marcel, car il était un héros, et les héros ne pénètrent jamais les secrets de leurs femmes.

Mais, tout en ignorant ces amours, Kraken en goûtait les précieux avantages.

Le cosinus entre ces deux phrases est de 0.74 ; elles sont sémantiquement très liées. Ce résultat n'est pas étonnant puisque 3 termes sont répétés dans ces deux phrases (ignorer, Kraken, amour). 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 termes en commun. Dans l'exemple qui suit, aucun terme n'est répété, mais certains d'entre eux sont sémantiquement proches (incendie/flammes-inflammable ; camp/tentes), ce qui donne un cosinus assez élevé entre les deux phrases (0.56). Cette proximité sémantique a été extraite par l'analyse sémantique latente de l'ensemble des cooccurrences dans le corpus, soit parce que par exemple les mots *camp* et *tente* cooccurrent dans les mêmes phrases, soit parce que chacun cooccurre avec les mêmes autres mots (*scouts*, *vacances*, ...).

Par exemple :

L'incendie s'est déclaré peu avant midi dans ce camp érigé dans la plaine de Mina, près de la ville sainte.

Les flammes se sont rapidement propagées, les tentes - souvent en tissu inflammable - se trouvant très près les unes des autres.

Extrait d'un corpus composé d'articles de journaux du Soir (97-98)

Utilisation des cosinus dans nos études

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. Par contre, 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.

Reprenons notre exemple :

p-1 Et plusieurs disaient, dans l' île, que, la nuit, sur les chemins passaient des anges et qu' on mourait pour les avoir vus.

p Kraken ignorait les amours d' Orberose et de Marcel, car il était un héros, et les héros ne pénètrent jamais les secrets de leurs femmes.

p+1 Mais, tout en ignorant ces amours, Kraken en goûtait les précieux avantages.

Le cosinus entre la phrase cible et la phrase qui précède (Cos [p-1,p] = 0.06) est plus petit qu'entre la phrase cible et la phrase qui la suit (Cos [p,p+1] = 0.74). La phrase cible est donc plus liée sémantiquement à la phrase qui la suit.

Il est à noter que les décisions prises dès la constitution du tableau lexical et tout au long d'une l'analyse sémantique latente peuvent en affecter les résultats. Il s'agit principalement de la sélection des documents (longueurs) et des termes (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. On notera néanmoins que Bestgen (2004) a montré que l'efficacité d'un algorithme de segmentation automatique de textes, basé sur l'ASL (Choi & al., 2001), n'était que peu influencée par ces paramètres.

Ces deux indices vont nous permettre d'étudier la fonction de marqueurs de la segmentation des expressions qui nous intéressent. Nous devons d'abord valider la technique qui permet de repérer automatiquement les expressions qui signalent la structure d'un texte. Ensuite, nous allons tester si elle arrive à discriminer la capacité de marquage de différents types d'expressions temporelles. Puis, nous allons comparer la fonction d'indicateur de la structure du discours d'expressions référentielles et d'adverbiaux temporels. Après avoir comparé ces deux types d'expressions, nous allons analyser ce qu'il en est de leur usage simultané.

Chapitre 2 : validation de la technique de recherche automatique de marqueurs de la segmentation

Ce chapitre est un remaniement de Piérard & Bestgen (2006a). « À la pêche aux marqueurs linguistiques dans la structure du discours ». Actes des 8èmes journées internationales d'analyse statistique des données textuelles (JADT06), Besançon, 19-21 avril 2006, Presses Universitaires de Franche-Comté, p.749-757.

Pour observer les expressions qui nous intéressent en corpus, il faut tout d'abord vérifier que notre méthode fonctionne. La première étude a pour objectif de tester une technique automatique de recherche d'expressions qui signalent la structure d'un texte. Un auteur signale les relations entre les phrases ou entre les segments de texte en intégrant des marqueurs tout au long de son discours ; c'est ce qu'on appelle les marqueurs globaux du discours. Ces marqueurs peuvent être des expressions temporelles ou spatiales, des expressions métadiscursives ou encore des expressions adverbiales ou référentielles (Bestgen & Costermans, 1994 ; Charolles, 1997 ; Lenk, 1998 ; Marcu, 2000, Segal & al. , 1991, Virtanen, 1992). Nous allons prouver l'efficacité de notre méthode en cherchant ces marqueurs de manière automatique. On sait qu'ils jalonnent les textes donc, si notre technique est efficace, elle doit pouvoir les mettre en évidence.

1. Méthodologie proposée

La procédure proposée est composée de deux étapes. La première consiste à récolter le plus grand nombre possible d'expressions qui pourraient fonctionner comme marqueurs. La seconde étape a pour fonction de filtrer cette vaste liste afin de ne conserver que les expressions qui fonctionnent effectivement comme marqueurs de la structure, et pour le cas qui nous occupe, les marqueurs de

discontinuité thématique. C'est l'efficacité de nos deux indices employés lors de cette seconde étape qui nous intéresse tout particulièrement.

Pour dresser la liste des marqueurs potentiels, nous avons opté pour l'identification automatique des séquences récurrentes de mots, aussi appelées *N-grams,* qui est fréquemment employée en phraséologie (Altenberg, 1998 ; Degand & Bestgen, 2003; Hernandez & Grau, 2003 ; Schone & Jurafsky, 2001), mais aussi en statistique textuelle (Lebart & Salem, 1992). Il s'agit de dresser la liste de tous les N-grams, N allant de 1 à 5 par exemple, suffisamment fréquents dans un corpus. La seule difficulté résultant de cette façon de procéder est que le nombre d'expressions récurrentes recueillies est très important. Une manière classique de limiter ce nombre est d'exiger une fréquence minimale dans le corpus. Cela permet d'éliminer immédiatement toute séquence dont un des mots à une fréquence totale inférieure à ce seuil.

La seconde étape est beaucoup plus complexe puisqu'elle doit permettre de filtrer cette liste afin de ne garder que les expressions qui présentent les qualités souhaitées. Pour ce faire, nous proposons de tirer parti de deux propriétés centrales des marqueurs de la structure du discours : ils apparaissent en début de phrase[6] et, de par leur statut de marqueurs de discontinuité, ils apparaissent en des lieux de rupture thématique. La première contrainte nous a conduits à n'analyser que les N-grams qui se trouvent en début de phrases. Afin de tenir compte de la deuxième contrainte, nous avons utilisé les deux indices présentés précédemment, à savoir le paragraphe et un indice de cohésion lexicale issu de l'analyse sémantique latente.

[6] Nous n'aborderons pas les cas où le marqueur est une phrase à lui seul (« Maintenant, je vais développer mon second argument »).

2. Objectif

L'objectif de la présente étude est d'évaluer la capacité des deux indices à extraire d'un corpus des expressions qui présentent les propriétés des marqueurs de la discontinuité thématique. La difficulté principale que rencontre cette entreprise est qu'on ne dispose pas d'une liste indépendante de ces expressions permettant de décider si la "pêche" a été bonne. Aussi, afin de comparer le plus objectivement possible ces deux indices, nous emploierons comme test une expérience de validation sur un corpus indépendant. Concrètement, les deux indices seront employés pour sélectionner des expressions dans un premier corpus, le corpus d'apprentissage, composé d'articles parus dans le journal « le Soir » en 1996. Dans un second temps, nous établirons si les expressions sélectionnées fonctionnent également comme des marqueurs de rupture dans un second corpus, le corpus de test, composés d'articles parus durant l'année suivante dans le même journal.

3. Le corpus

Nous avons testé l'efficacité des deux indices de sélection des marqueurs sur la base d'un corpus de 37 000 articles parus en 1996 dans le journal belge francophone *Le Soir*. Dans un premier temps, une série d'articles ont été supprimés parce que leur contenu était peu propice aux objectifs de cette étude : résultats sportifs, info-routes, etc. Comme on peut penser que les marqueurs de la structure sont différents à l'oral et à l'écrit, nous avons retiré du corpus les retranscriptions de discours oraux. Ces passages étant formatés en italique, il a été possible de les identifier d'une manière automatique. Ceci nous a permis de supprimer tous les textes qui contenaient une proportion importante de passage en italique et un nombre suffisamment élevé de basculements entre le format normal et l'italique. À l'issue de ces traitements, le corpus était composé d'approximativement 13 000 000 de mots.

Dans un second temps, le corpus a été lemmatisé au moyen du programme TreeTagger de Schmid (1994). C'est également TreeTagger qui a segmenté les articles en phrases. L'identification des paragraphes n'a pas posé de problèmes puisque ceux-ci sont codés dans le corpus par deux retours à la ligne successifs.

L'indice de cohésion lexicale nécessite, pour être calculé, un espace sémantique obtenu par décomposition en valeurs singulières d'un tableau lexical. Nous avons construit ce tableau sur la base du corpus d'articles de journaux décrits ci-dessus. Ce corpus a été segmenté en articles et tous les mots mentionnés au moins deux fois, mais absents d'une liste de mots fonctionnels, 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 pour mesurer les proximités entre les phrases.

Afin de disposer d'un corpus de validation, nous avons traité exactement de la même manière les articles parus dans *Le Soir* en 1997. Un espace sémantique, répondant aux mêmes critères, a été dérivé.

4. Résultats

4.1 Validation de l'indice de cohésion

Dans un premier temps, nous avons effectué une série d'analyses afin de confirmer que l'indice de cohésion lexicale permettait bien de détecter des ruptures thématiques. Plus précisément, nous avons calculé la différence des cosinus pour la première phrase de chaque article. On peut en effet espérer que celle-ci est plus liée à celle qui la suit qu'à la dernière phrase de l'article précédent[7]. On peut aussi s'attendre à ce que la différence des cosinus pour la

[7] Notons toutefois que les articles sont ordonnés dans le corpus en fonction de leur date de parution et de leur position dans le journal. Il s'ensuit que les articles thématiquement liés (politique intérieure, culture, sport) se suivent dans le corpus.

première phrase d'un paragraphe soit plus grande que la différence des cosinus pour les phrases internes à un paragraphe. Ces prédictions sont en effet vérifiées, puisque la différence des cosinus pour la première phrase d'un article est de 0.13 alors qu'elle n'est que de 0.03 pour les premières phrases des paragraphes et que de -0.02 pour les phrases internes aux paragraphes. Vu le nombre de cas de chaque type (entre 36 000 et 380 000), ces valeurs sont évidemment significativement différentes pour une analyse de la variance.

4.2 Procédure d'analyse statistique

Les résultats se présentent sous la forme suivante. Pour chaque N-gram sélectionné, on dispose du nombre de fois où il apparaît en tête d'un paragraphe et du nombre de fois où il est en milieu de paragraphe. En prenant en compte le nombre total de paragraphes et le nombre total de phrases, il est possible de construire la table de contingence donnée ci-dessous pour le 3-gram *À l'étranger*. Sur la base de ce type de tables, nous avons calculé le rapport des chances (RC) pour un N-gram donné d'être présent en début de paragraphe. Pour l'exemple du tableau 24, il apparaît qu'une phrase qui commence par le 3-gram *A l'étranger* a 11.82 fois plus de chances d'être en tête de paragraphe qu'en milieu (RC = [14 / 4] / [102 766 / 346 932]). Nous avons préféré cette statistique à celle du Chi² parce que contrairement à ce dernier, elle n'est pas influencée par des totaux inégaux de ligne et donc par le fait que certains N-grams sont nettement plus fréquents que d'autres (Howell, 1998, p. 182). Nous avons néanmoins employé le Chi² pour éliminer tous les N-grams qui n'étaient pas significativement plus fréquents en début de paragraphe qu'au milieu ($p \leq 0.05$). Nous avons aussi éliminé les N-grams qui apparaissaient moins de 5 fois en tête de paragraphe dans le corpus.

Tableau 24 : table de contingence pour l'expression *A l'étranger*

		Début de paragraphe		
		OUI	NON	Total
N-gram présent	OUI	14	4	18
	NON	102 766	346 932	449 698
	Total	102 780	346 936	449716

Les mêmes analyses ont été effectuées pour l'indice de cohésion. Il faut toutefois noter que, contrairement à l'indice *paragraphe*, la différence entre les cosinus est une grandeur continue. Au risque de perdre un peu de puissance, nous avons choisi, dans cette étude exploratoire, de le traiter comme une variable dichotomique afin de pouvoir lui appliquer les mêmes statistiques que celles appliquées à l'indice basé sur les paragraphes. Une phrase est considérée comme discontinue par rapport à celle qui la précède lorsque la différence entre les deux cosinus est supérieure à 0.

4.3 Résultats

La figure 19 présente le nombre de N-grams sélectionnés par les indices en fonction de leur rapport des chances (RC) calculé tel qu'expliqué ci-dessus. Les N-grams vont de 1 à 5. Si un N-gram de longueur plus petite (comme le 2-gram *A l'*) est inclus dans un N-gram plus long (comme le 3-gram *A l'étranger*), le 2-gram n'est pas pris en compte au profit du 3-gram. Il s'agit d'une distribution cumulée, un RC de 10 correspondant à tous les RC supérieurs ou égaux à cette valeur, un RC de 9 à tous ceux qui sont supérieurs ou égaux à cette valeur et ainsi de suite.

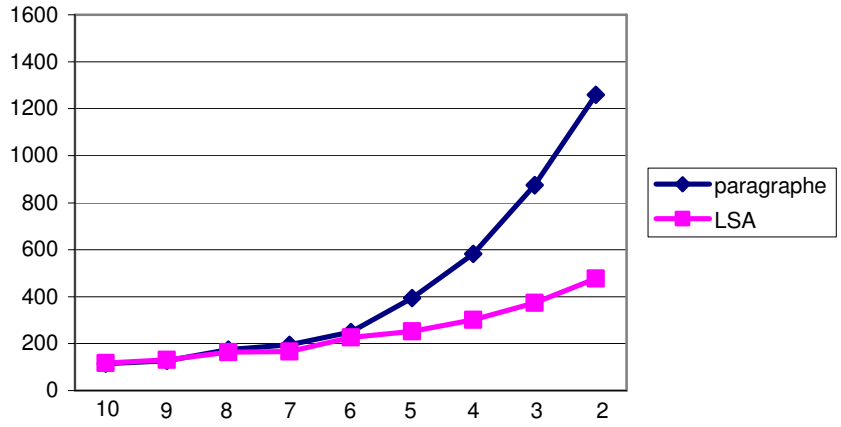


Figure 19 : nombre de N-grams sélectionnés selon le RC

Comme on peut le voir, jusqu'à un RC plus grand ou égal à 7, les deux indices sélectionnent le même nombre d'expressions. Au-delà, l'indice basé sur le paragraphe permet la sélection d'un plus grand nombre d'expressions.

La figure 20 présente le même genre de données, mais pour le nombre de phrases contenant ces expressions qui sont soit en début de paragraphe, soit en situation de discontinuité pour l'indice lexical (différence des cosinus positive). Les résultats sont très similaires.

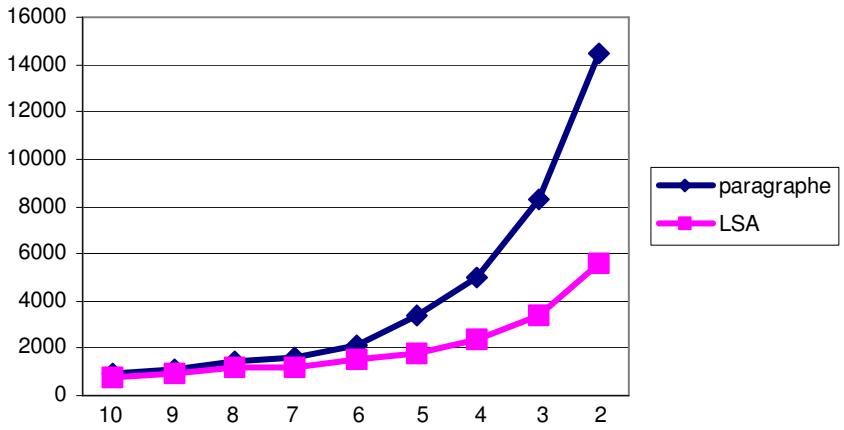


Figure 20 : nombre de phrases contenant les N-grams sélectionnés

L'analyse la plus intéressante porte sur le test de validation appliqué à l'année suivante du même journal. Nous avons procédé en recherchant dans ce second corpus tous les N-grams sélectionnés dans le corpus de 1996, sans plus appliquer de seuil de fréquence minimale, et nous avons déterminé si les phrases qui contenaient ces N-grams étaient en situation de discontinuité, c'est-à-dire dans la première phrase d'un paragraphe lorsqu'elles avaient été sélectionnées par cet indice ou précédés par une différence de cosinus positive lorsqu'elles avaient été sélectionnées par l'indice "cohésion". Les résultats sont présentés dans la figure 21.

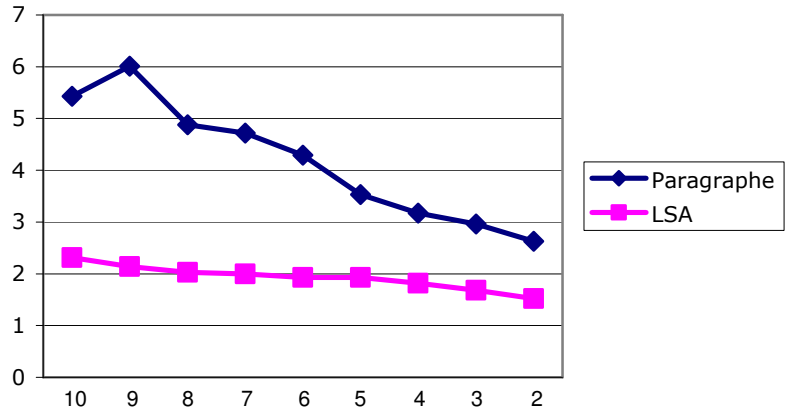


Figure 21 : rapports des chances moyens des expressions sélectionnées dans le corpus de test

On y a représenté les rapports des chances moyens calculés dans le corpus de test pour les expressions sélectionnées dans le corpus d'apprentissage, et ce, en fonction du rapport des chances dans ce corpus d'apprentissage. En d'autres mots, la valeur de 5.40 pour le 10 en abscisse correspond au rapport des chances moyen pour les N-grams sélectionnés par l'indice de paragraphe qui, dans le corpus d'apprentissage, ont un rapport des chances supérieur ou égal à 10.

On observe que les N-grams sélectionnés par l'indice *paragraphe* sont nettement plus fréquemment associés à des ruptures de continuités dans ce second corpus que les N-grams sélectionnés par l'indice *de cohésion*.

5. Conclusions

Nous avons proposé deux indices pour identifier dans un corpus de textes des marqueurs de structure textuelle. La technique fonctionne : les N-grams trouvés lors de la première étape sont bien associés à des ruptures de thème dans le corpus de test. On observe que les N-grams sélectionnés par l'indice *paragraphe* sont nettement plus fréquemment associés à des ruptures thématiques dans ce second corpus que les N-grams sélectionnés par l'indice *de cohésion*. Cependant, ces deux indices sont complémentaires. L'indice de cohésion lexicale a un caractère local. Il indique qu'une phrase est sémantiquement plus liée avec celle qui la suit qu'avec celle qui la précède. Le paragraphe est par contre une mesure plus globale puisqu'il segmente le texte en paquets de phrases qui forment à chaque fois une unité textuelle. Pour cette raison, nous continuerons d'employer ces deux indices dans les recherches suivantes.

Nos deux indices permettent effectivement de trouver des marqueurs de la structure d'un texte ; nous allons maintenant expérimenter notre technique sur un type de marqueurs précis : les marqueurs temporels.

Chapitre 3 : différence de capacité de marquage des expressions temporelles et comparaison d'indices

Ce chapitre est un remaniement de Piérard, Degand & Bestgen (2004). « Vers une recherche automatique des marqueurs de la segmentation du discours ». In G. Purnelle, C. Fairon & A. Dister (Eds). Actes des 7èmes journées internationales d'analyse statistique des données textuelles (JADT04), Louvain-la-Neuve, 10-12 mars 2004, Presses Universitaires de Louvain, p.171-181.

Cette étude de corpus cible une partie des marqueurs qui nous intéressent : les marqueurs temporels. Un de leurs avantages est qu'ils forment une échelle allant du signalement d'une rupture importante à celui d'une continuité importante. Afin de montrer si la validité de cette échelle peut être confirmée par la technique basée sur nos indices, nous avons comparé l'indice de cohésion à trois autres indices qu'il est possible de calculer, au moins d'une manière approximative, par des procédures automatiques.

1. Marqueurs étudiés

Nous avons sélectionné un ensemble d'expressions temporelles dont l'importance a été soulignée par de nombreux auteurs (p.e. Segal & al., 1991 ; Virtanen, 1992). Les marqueurs d'ancrage (*vers dix heures*) qui font référence à un temps absolu, extérieur à la narration, signalent les ruptures les plus importantes. Les marqueurs d'enchaînement (*puis*, *ensuite*) qui indiquent la succession des actions à l'intérieur d'une narration signalent les ruptures intermédiaires alors que le connecteur *et* souligne la continuité discursive. Introduits en début de phrase, ces dispositifs peuvent être employés pour signaler la structure hiérarchique d'une narration comme l'a montré l'analyse d'un petit corpus de narrations de journées (Costermans & Bestgen, 1991).

2. Indices de continuité/discontinuité

L'objectif de cette étude est de démontrer de manière automatique que les marqueurs temporels sont associés à des degrés de marquage différents. Pour ce faire, nous avons employé l'indice de cohésion lexicale (dit « cosinus ») que nous avons comparé à trois autres : un indice référentiel, un indice anaphorique et un indice obtenu à partir de l'algorithme de segmentation de Choi & al. (2001) (dit « segment »).

Les deux premiers indices s'inscrivent dans la ligne des travaux de Passoneau & Litman (1997) et Beeferman, Berger & Lafferty (1999) qui ont proposé de détecter les changements de thème dans un texte sur la base, entre autres, des ruptures dans les chaînes référentielles. Le premier indice est une mesure référentielle inspirée du modèle de Kintsch & van Dijk (1978). Ces auteurs ont élaboré un système de représentation de la signification d'un texte au moyen d'une liste structurée de propositions composées d'un prédicat et d'un ou de plusieurs arguments. Dans ce modèle, deux phrases sont liées si les propositions qui les composent contiennent des arguments communs. En suivant Costermans (1998), nous définissons les arguments comme étant des noms ou des pronoms. Nous avons donc identifié dans chaque phrase cible, c'est-à-dire qui contient un marqueur, le nombre d'arguments présents dans au moins une des cinq phrases qui la précède ainsi que le nombre d'arguments neufs. S'il y a continuité thématique, le nombre d'arguments communs devrait être important. Dans le cas d'une rupture, les arguments neufs devraient être plus nombreux. Cet indice a été calculé par un juge et par une procédure automatique basée sur la comparaison des noms et pronoms présents dans la phrase cible et dans celles qui la précèdent. La limite majeure de cette procédure automatique est que deux arguments peuvent faire référence à la même entité en prenant des formes différentes comme *commerçant/marchand* ou encore *Robert/il.*

La deuxième mesure ne prend en compte que les anaphores grammaticales. Selon l'expression de Moeschler « l'anaphore est une expression référentielle non autonome ». Sa présence est donc la trace d'une continuité thématique. Nous avons donc comptabilisé dans chaque phrase cible l'ensemble des pronoms et syntagmes nominaux anaphoriques (personnel, possessif et démonstratif) en distinguant les anaphores intra-phrase (*Léon fit démarrer sa voiture*) des anaphores inter-phrases (*Léon devait aller à l'aéroport. Sa voiture était en panne*) seules pertinentes ici. L'automatisation de cette procédure a simplement consisté en l'identification dans la phrase cible de l'ensemble des pronoms et déterminants pouvant être anaphoriques. Les problèmes majeurs de cette implémentation résident en la distinction entre les anaphores intra et inter-phrases et du fait qu'un terme peut endosser plusieurs fonctions grammaticales : *le son de la cornemuse/ son chien*.

Le troisième indice (Segment) est obtenu à partir de l'algorithme de segmentation de Choi & al. (2001). Celui-ci a été paramétré pour identifier d'une manière récursive les ruptures entre les 10 phrases d'un extrait en allant de la plus importante à la moins importante. Une rupture est donc d'autant plus forte qu'elle a été identifiée parmi les premières. L'indice employé est l'ordre d'importance de la segmentation qui précède la phrase cible en valeur relative, c'est-à-dire divisée par le nombre maximum de ruptures (9). Comme pour les autres indices, une valeur élevée indique à chaque fois une forte continuité.

3. Expérience

L'objectif de cette expérience est de tester l'efficacité des indices de continuité/discontinuité décrits ci-dessus pour l'étude du fonctionnement des marqueurs de la segmentation en montrant que les marqueurs d'ancrage, d'enchaînage et le connecteur *et* sont associés à des niveaux de continuité/discontinuité thématiques différents. Pour deux de ces indices, nous

comparons également les valeurs obtenues par une procédure automatique à celles déterminées par un juge.

3.1 Corpus et extraction des marqueurs en contexte

Le corpus est composé de textes littéraires (romans, nouvelles et contes) extraits principalement des bases ABU[8] et Frantext. Il contient 5 000 000 de mots. Les textes ont été découpés en phrases et lemmatisés au moyen du programme *TreeTagger* de Schmid (1994). Pour en extraire les expressions appartenant aux trois catégories de marqueurs, nous avons recherché toutes les occurrences en début de phrases des formes suivantes[9].

— « A midi », « Vers midi », « A minuit », « Vers minuit », « A *x* heure(s) », « Vers *x* heure(s) » où *x* représente un nombre compris entre un et vingt-quatre.

— Puis, Ensuite

— Et

Sur cette base, on a procédé à une sélection aléatoire d'un échantillon de 90 phrases pour chacune des trois catégories de marqueurs. Cette sélection a été effectuée en prenant en compte les oeuvres littéraires et les auteurs de telle sorte qu'un nombre équivalent de phrases appartenant à chaque condition soit extrait d'un auteur donné. Enfin, on a extrait le contexte qui entoure les phrases contenant les marqueurs, soit les cinq phrases qui précèdent la phrase cible et les quatre phrases qui suivent celle-ci.

---

[8] http://abu.cnam.fr/

[9] Afin de se focaliser sur l'emploi de ces expressions à l'écrit, nous n'avons pas pris en compte les expressions apparaissent dans un dialogue. Ceux-ci ont cependant pas été retirés du corpus afin d'optimaliser le calcul de l'indice « segment ».

Pour l'estimation de la continuité/discontinuité par les chaînes référentielles, les deux indices suivants ont été calculés[10] :

— Indice référentiel : le nombre d'éléments liés divisé par la somme du nombre d'éléments liés et d'éléments non liés.

— Indice anaphorique : le nombre total d'anaphores identifiées dans la phrase cible.

Pour les analyses manuelles de ces deux indices, un sous-échantillon de plus ou moins 50 segments par catégorie de marqueurs a été extrait. Dans chacun de ces segments, les marqueurs ont été supprimés par une personne qui n'a pas participé au codage manuel des phrases cibles. Un premier juge a calculé pour chaque extrait les deux indices de continuité définis ci-dessus. Afin d'estimer la fiabilité de cette évaluation, un second juge a analysé indépendamment 12 extraits de chaque type de marqueurs. Pour chacun des deux indices, la corrélation entre les évaluations des deux juges est au moins égale à 0.79.

Pour les analyses basées sur l'analyse sémantique latente, nous avons extrait du corpus initial un espace sémantique en respectant les paramètres proposés par Bestgen (2004). Le premier indice (Cosinus) employé est basé sur les cosinus entre une phrase cible (p) et celles qui la précède (p-1) et qui la suit (p+1). Si une phrase introduit un changement de thème, son cosinus avec celle qui la précède devrait être plus petit que celui avec la phrase qui la suit. Nous utiliserons donc comme indice : cos(p,p+1) - cos(p-1,p). Le second indice (Segment) est obtenu à partir de l'algorithme de segmentation de Choi & al. (2001) comme expliqué précédemment.

[10] Ces indices sont des mesures.

3.2 Résultats

Comme indiqué ci-dessus, la procédure manuelle n'a été appliquée qu'au sous-échantillon de 152 phrases cibles. La procédure automatique a, par contre, été appliquée à ce sous-échantillon, mais aussi à l'échantillon total de 270 phrases cibles. Pour évaluer la fiabilité des analyses automatiques, nous avons corrélé les indices obtenus par ces procédures avec les évaluations faites par le juge. Tant pour l'indice référentiel que pour l'indice anaphorique, la corrélation entre est de 0.59 ($p<0.0001$). Au vu du caractère très rudimentaire de la procédure d'analyse automatique, ce résultat est très encourageant et valide l'utilisation de techniques automatiques dans l'étude de grands corpus de textes. Comme il n'y a pas de procédure manuelle pour calculer les indices obtenus à partir de l'ASL, ceux-ci n'ont été calculés que pour l'échantillon complet.

Le tableau 25 présente les valeurs moyennes de continuité pour les trois catégories de marqueurs. La présence de différences statistiquement significatives a été déterminée par des analyses de variance à un critère.

Tableau 25 : Valeurs moyennes de continuité pour les trois catégories de marqueurs

	Indice référentiel			Indice anaphorique			ASL (Auto.)	
	Manuel** N=152	Auto* N=152	Auto** N=270	Manuel*** N=152	Auto*** N=152	Auto*** N=270	Cosinus* N=270	Segment** N=270
Et	0.44	0.39	0.35	1.00	1.63	1.60	+0.03	0.57
ENC	0.34	0.33	0.30	0.94	1.37	1.40	+0.02	0.55
ANC	0.23	0.26	0.24	0.22	0.61	0.60	-0.05	0.43

Légende : Une valeur élevée indique une continuité forte. * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Globalement, les valeurs observées sont ordonnées comme attendu. Le connecteur *et* est associé à la plus grande continuité ; le marqueur d'ancrage à la

plus forte discontinuité. Toutefois, pour trois des quatre mesures, le marqueur d'enchaînement est nettement plus proche de *et* que de l'ancreur.

4. Conclusions

Nous avons pu analyser dans un grand corpus de textes les fonctions de marqueurs de la continuité/discontinuité d'un ensemble d'expressions temporelles. Si les résultats confirment globalement l'échelle postulée, ils suggèrent la nécessité d'une analyse plus profonde de la distinction entre les enchaîneurs et le connecteur *et*.

Au niveau des techniques de segmentation automatique des textes, un résultat mérite d'être mis en exergue. L'indice anaphorique a donné lieu à des différences entre les trois catégories de marqueurs beaucoup plus significatives que celles obtenues avec les autres indices. Or, cet indice est rarement pris en compte par les algorithmes de segmentation automatique, qui, classiquement, commencent par supprimer les mots les plus fréquents et les mots outils. S'il est exact que ce genre de termes pose problème pour l'analyse de la cohésion lexicale, il serait intéressant de développer des approches qui combinent la cohésion lexicale et grammaticale.

Les anaphores semblent fonctionner relativement bien pour indiquer la structure d'un texte. Elles seront l'objet d'étude des deux expériences suivantes.

Chapitre 4 : comparaison de la fonction d'indicateur de la structure d'expressions référentielles et temporelles

Ce chapitre est un remaniement de Piérard & Bestgen (à paraître). « Deux indices pour l'étude des marqueurs de la continuité thématique dans de grands corpus ». Actes des 4èmes journées de linguistique de corpus, Lorient, 15-17 septembre 2005.

Nous savons que notre technique repère les marqueurs de la structure. Nous avons confirmé de manière automatique que les différents types de marqueurs temporels sont associés à des niveaux de continuité/discontinuité différents. Dans la troisième étude, nous allons comparer la fonction d'indicateur de la structure du discours des éléments linguistiques qui, selon la littérature, devraient fonctionner comme des signaux de la structure d'un texte.

1. Éléments linguistiques analysés

Il s'agit tant d'éléments censés signaler des ruptures thématiques comme les adverbiaux temporels cadratifs (« le matin») ou les groupes nominaux introduits par un article indéfini (« un … », « des … ») que d'éléments censés signaler une continuité thématique tels les pronoms personnels (« il », « elles »), les déterminants possessifs (« son », « ses », « sa ») et démonstratifs (« cet », « ces ») et les groupes nominaux introduits par un article défini (« le… », « l'… »). Linguistiquement, ces marques de continuité et de rupture sont unies par leur propriété référentielle très largement non autonome. Étudier simultanément ces deux types d'éléments est une des originalités de cette expérience. Jusqu'à présent en effet, les recherches, effectuées sur de petits corpus, ont principalement porté sur des dispositifs marquant une rupture comme les expressions adverbiales temporelles et spatiales (Charolles, 1997 ; Virtanen, 1992). Si les marques qui signalent la continuité discursive ont été largement

étudiées en linguistique, peu de travaux ont évalué leur fonction de signaux du discours et aucun, à notre connaissance, ne s'est intéressé à l'expression « déterminant possessif + syntagme nominal » (SN possessif). La spécificité de cette dernière est qu'elle est biréférentielle : elle représente un référent par rapport à un autre (Apothéloz, 1995 ; Broeder, 1992). Bien que le syntagme nominal soit l'élément principal, le déterminant possessif devrait souligner la continuité avec les phrases précédentes (Piérard & Bestgen, 2005). Analyser cette expression avec les déterminants démonstratifs (Botley & McEnery, 2001) et les pronoms personnels (Delisle, 1993), considérés comme des marques de continuité dans un texte, sera particulièrement instructif. Nous allons pouvoir en effet comparer leurs efficacités.

2. Corpus

Nous avons travaillé sur un corpus composé de textes littéraires. Nous n'avons pas gardé le même corpus que dans l'expérience précédente. En effet, dans ce dernier, les œuvres de Maupassant étaient sur représentées[11]. Nous avons donc veillé dans notre nouveau corpus, à ne prendre que maximum deux romans du même auteur. Ceux-ci étaient extraits des bases ABU[12], Intratext[13] et Wordthèque[14]. Ce corpus contient 67 romans du XIXème et XXème 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

[11] Nous avons tenu compte de ce paramètre dans l'analyse de corpus 2 et nous avons veillé à ce qu'un nombre équivalent de phrases appartenant à chaque condition soit extrait d'un auteur donné.

[12] http://abu.cnam.fr/

[13] http://www.intratext.com/Aiuto/FRA/

[14] http://www.logoslibrary.eu

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, le corpus contient approximativement 107 000 phrases et 2 270 000 mots.

Ce même corpus, dans sa version lemmatisée et expurgée des dialogues et des formes fonctionnelles (pronoms, articles, ...), a été employé 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 le corpus 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.

Le tableau 26 donne la liste des éléments linguistiques étudiés et le nombre de ceux-ci apparaissant au tout début d'une phrase. Comme adverbiaux cadratifs, nous avons employé les expressions temporelles incluant une référence à une des grandes subdivisions de la journée (*matin, matinée, après-midi, soir, soirée, nuit*) dont Costermans & Bestgen (1991) ont montré la fonction de marqueurs de discontinuité thématique. En ce qui concerne les pronoms personnels et les SN possessifs, nous n'avons pris en compte que la troisième personne.

Tableau 26 : fréquence dans le corpus des expressions linguistiques analysées

	Fréquence
Adverbiaux cadratifs temporels (CAD)	581
Articles définis (DEF)	10343
Articles indéfinis (IND)	2167
SN démonstratifs (DEM)	1792
SN possessifs 3e pers. (POS)	1478
Pronoms personnels 3e pers. (PPE).	12915

Ce même corpus, dans sa version lemmatisée et expurgée des dialogues, a été employé pour construire l'espace sémantique nécessaire pour l'analyse sémantique latente.

3. Résultats

3.1 Analyse de la position dans les paragraphes

La figure 22 présente l'ensemble des résultats obtenus avec la mesure du Chi² et du rapport des chances. L'axe gradué indique dans quelle proportion une expression est observée plus souvent en début de paragraphe qu'au milieu (côté droit - rapport supérieur à 1) ou l'inverse (côté gauche - rapport inférieur à 1). Par exemple, la valeur "3/1" indique qu'un élément a trois fois plus de chance d'apparaître en début de paragraphe qu'au milieu ; la valeur "1/2" indique qu'un élément a deux fois plus de chance d'apparaître au milieu d'un paragraphe qu'au début, la valeur "1/1" signalant un élément apparaissant aussi souvent au début qu'au milieu. Les tests du Chi² indiquent que tous les éléments linguistiques analysés ont un rapport des chances significativement différent de 1/1. Selon cet indice, ils sont donc tous des indicateurs de continuité ou de rupture.

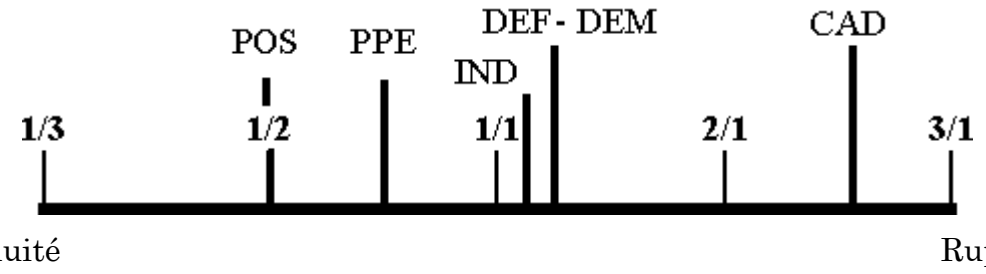


Figure 22 : distribution des expressions en fonction du rapport des chances

Globalement, ces résultats correspondent à nos attentes. Les adverbiaux cadratifs temporels sont bien des marques de discontinuité et les pronoms personnels des marques de continuité. L'article indéfini est bien du côté

discontinu de l'échelle. On note toutefois qu'il est très proche du point neutre, même s'il en est significativement différent selon le test du Chi². On note surtout que l'article défini et le SN démonstratif s'observent plus fréquemment que lui en début de paragraphe. Ce résultat, moins attendu, sera discuté après l'analyse du second indice. Enfin, le SN possessif fonctionne comme les pronoms.

3.2 Analyse de la proximité thématique

La figure 23 présente les résultats obtenus par l'analyse de la proximité thématique. L'axe gradué indique la valeur du cosinus. Plus cette valeur est élevée, plus les phrases analysées sont thématiquement proches. La partie supérieure de la figure présente les cosinus moyens entre la phrase qui commence par l'élément en question et la phrase qui la précède, la partie inférieure rapportant les cosinus moyens entre cette même phrase cible et celle qui la suit.

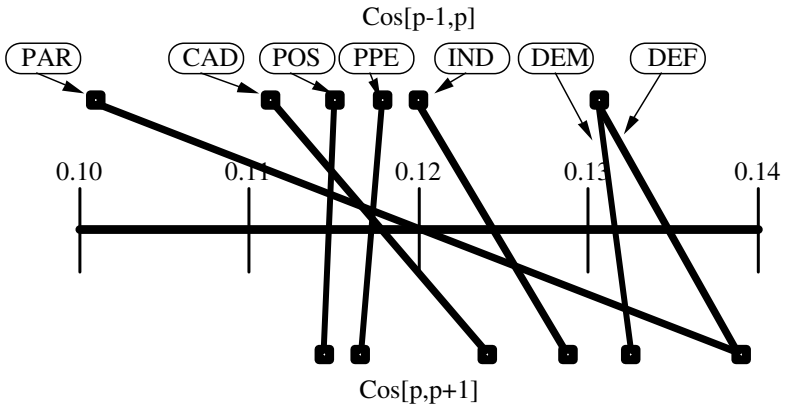


Figure 23 : distribution des expressions en fonction des cosinus

À titre d'exemple, nous avons calculé les cosinus moyens pour la première phrase d'un paragraphe (PAR). Comme attendu, on observe un cosinus moyen plus faible (0.10) entre la première phrase d'un paragraphe et celle qui la précède qu'entre cette même phrase et celle qui la suit (0.14), confirmant la fonction de marqueur de rupture du changement de paragraphe.

Les adverbiaux cadratifs et les articles indéfinis présentent également le profil attendu de marques de discontinuité. Les pronoms personnels et les SN possessifs se comportent plutôt comme des marques de continuité, confirmant les résultats de la première analyse. Les SN démonstratifs et les articles définis méritent une attention particulière. Ils présentent globalement des cosinus très élevés tant avec la phrase qui précède qu'avec celle qui suit. Cette observation permet, selon nous, de mieux comprendre les résultats obtenus lors de l'analyse des changements de paragraphe. On peut en effet penser qu'un certain nombre de changements de paragraphe n'ont pas comme objectif de signaler une rupture thématique, mais plutôt de mettre en évidence le nom qu'ils déterminent (Stark, 1988). Une autre interprétation est que ces éléments sont employés pour souligner la continuité entre deux paragraphes successifs et tout particulièrement le fait que ceux-ci portent sur un thème commun (Hofmann, 1989). Le cas des articles définis est encore plus particulier. Ceux-ci sont encore plus liés avec la phrase qui suit qu'avec celle qui précède, à la manière d'une marque de discontinuité, mais leur cosinus très élevé avec la phrase qui précède rend une telle interprétation difficile. Plus généralement, le fait que les phrases commençant par un article défini, un SN démonstratif ou même un article indéfini donnent lieu à des cosinus moyens plus élevés que ceux observés pour les pronoms personnels et les SN possessifs mérite des investigations complémentaires comme une analyse comparative fine d'échantillons de ces expressions.

4. Conclusions

Les résultats obtenus grâce aux deux indices sont globalement en accord avec la littérature : les éléments linguistiques censés indiquer un changement de thème sont bien repérés comme tels. Il en est de même pour les expressions signalant des continuités. Ces résultats autorisent une analyse plus fine des résultats et, plus particulièrement, la comparaison des différentes expressions entre elles. Ainsi, le déterminant possessif apparaît moins souvent en début de paragraphe que le pronom personnel, soulignant sa fonction de signal de continuité, en accord

avec les analyses linguistiques montrant que l'antécédent du déterminant possessif est presque toujours au plus loin dans la phrase qui précède (Apothéloz, 1995).

La complémentarité des deux procédures est mise en évidence par l'interprétation du démonstratif et du déterminant défini comme des dispositifs signalant un glissement thématique : l'introduction d'un nouveau thème lié au précédent. Les SN démonstratifs et les articles définis présentent globalement des cosinus très élevés tant avec la phrase qui précède qu'avec celle qui suit. On peut penser qu'un certain nombre de changements de paragraphe n'ont pas comme objectif de signaler une rupture thématique, mais plutôt de mettre en évidence le nom qu'ils déterminent, de remettre dans le focus une entité jusque-là peu ou pas saillante (Stark, 1988 ; Charolles, 1994). Une autre interprétation des démonstratifs est qu'ils sont employés pour souligner la continuité entre deux paragraphes successifs et tout particulièrement le fait que ceux-ci portent sur un thème commun (Hofmann, 1989). Le cas des articles définis est encore plus particulier. Ceux-ci sont plus liés avec la phrase qui suit qu'avec celle qui précède, à la manière d'une marque de discontinuité, mais leur cosinus très élevé avec la phrase qui précède suggère une proximité thématique élevée entre ces phrases.

Ces interprétations ne remettent pas en cause celles avancées par Vonk & al. (1992) à propos des expressions surspécifiques. En effet, ces auteurs soutiennent l'idée qu'un changement de thème est signalé par une expression plus spécifique que nécessaire. Dans le cas qui précède, les phrases commençant par un syntagme nominal précédé d'un article défini ont un cosinus élevé avec la phrase qui précède, mais encore plus élevé avec celle qui suit : il reste néanmoins un indice de changement. De plus, Vonk & al. (1992) ont travaillé avec des noms propres principalement. Nous allons voir dans l'expérience qui suit, si leur usage est simultané avec un autre type de marqueurs et si cette accumulation indique des ruptures plus importantes dans les textes.

Chapitre 5 : emploi simultané ou exclusif des expressions temporelles et référentielles comme marqueurs de la segmentation

Ce chapitre est un remaniement de Piérard & Bestgen (2006b). « Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes ». In M.-P. Pery-Woodley & D. Scott (Eds), Traitement Automatique des Langues : Discours et Document : traitements automatiques, Vol. 47 (2).

Nous avons classé différents éléments en fonction de leur capacité à signaler des continuités/discontinuités thématiques. Qu'en est-il de leur emploi simultané ? Comme nous l'avons dit au début de cette partie, Vonk & al. (1992) affirment qu'il n'y a pas d'emploi simultané des expressions référentielles et des expressions temporelles pour indiquer un changement de thème. Un des objectifs de cette quatrième étude 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. Pour arriver à mettre cela en évidence, nous avons adopté une démarche progressive dont les étapes seront détaillées dans la suite. Nous avons extrait de manière automatique les expressions temporelles de notre corpus (le même que celui employé pour l'analyse de corpus précédente) ; nous nous sommes intéressés au positionnement des phrases contenant ces expressions temporelles et avons sélectionné les expressions temporelles les plus efficaces à savoir celles qui arrivent en tête de phrase et en tête de paragraphe. Ensuite, nous nous sommes intéressés aux expressions référentielles présentes dans les phrases qui contiennent ces expressions temporelles efficaces.

1. Expressions temporelles

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 27). Au total, les phrases sélectionnées représentent 3 % des phrases du corpus.

Tableau 27 : catégories des marqueurs temporels étudiés

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[15]	en été au printemps	42
tard/après[16]	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

[15] En raison du petit nombre de cas de cette catégorie dans le corpus, nous l'avons éliminée dans la suite de nos analyses.

[16] Les expressions de ces deux types étant très similaires, elles ont été regroupées dans la même catégorie.

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 avons employé l'indice issu de l'analyse sémantique latente au moyen duquel nous calculons une différence de proximité sémantique entre les phrases. D'une manière générale, les analyses qui suivent partent de l'observation d'un fait général pour se concentrer sur des phénomènes linguistiques de plus en plus précis.

1.1 Indice paragraphe

Nos analyses, illustrées dans la figure 24, 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.

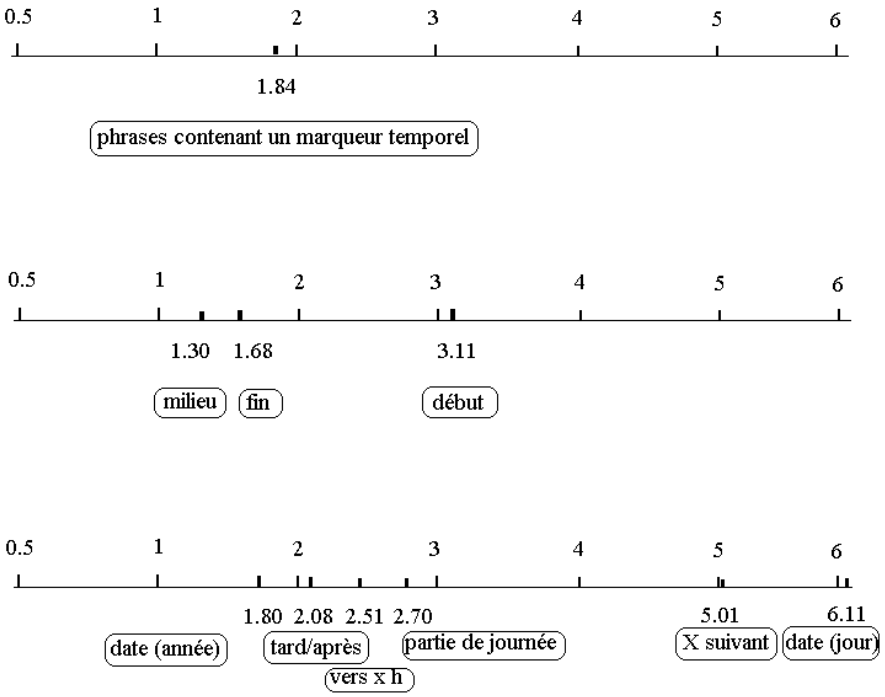


Figure 24 : rapport des chances des phrases contenant un marqueur temporel, en fonction de la place de celui-ci et selon la catégorie des marqueurs

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 & Bestgen, 1991 ; Virtanen, 1992).

Selon Costermans & Bestgen (1991), Segal & 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 24. Deux types d'expressions obtiennent des RC très élevés : les expressions exprimant une date (jour) et celles reprises dans la catégorie « X suivant ». Les phrases débutant par ces expressions sont celles qui, dans notre corpus, ont le plus de chance d'apparaître en tête de paragraphe. Ces expressions sont donc les plus efficaces pour indiquer un changement de thème.

1.2 Différence de cosinus

Les mêmes analyses que celles décrites ci-dessus ont été effectuées sur la base du second indice, la différence de 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 25 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. Par contre, 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 ($F(1,83533)= 1371.14$, $p<0.0001$). 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.

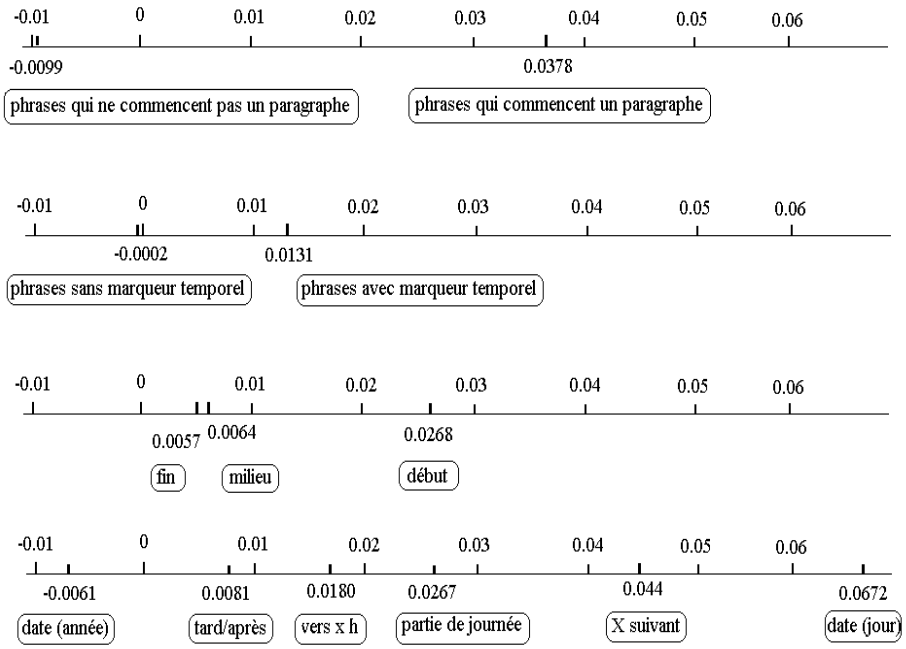


Figure 25 : différence de cosinus

Enfin, les expressions temporelles de début de phrases ont été classées en six catégories. À nouveau, les différences sont significatives ($F(5,857)=2.45$, $p<0.05$). La comparaison des figures 24 et 25 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 corpus avant de tenter de l'expliquer.

L'indice issu de l'ASL semble globalement prometteur. 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 conduits à ne plus employer que le seul indice basé sur les paragraphes.

2. Expressions temporelles et expressions référentielles

Comme déjà annoncé dans nos hypothèses, un objectif de cette étude est de mettre en évidence un éventuel usage simultané des expressions temporelles et 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 le 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 (SN possessif), Pronom personnel, Déterminant défini + SN (SN défini), Déterminant indéfini + SN (SN indéfini), Déterminant démonstratif + SN (SN démonstratif), 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 1028 phrases introduites par une expression temporelle, les seules analysées dans la suite[17], donne un pourcentage d'erreurs inférieur à 5 %, dont la moitié résulte d'un mauvais étiquetage par le 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 28 indique que le nom propre obtient un RC très élevé.

[17] 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.

Tableau 28 : RC des phrases débutant par un marqueur temporel selon le sujet grammatical (SN signifie syntagme nominal)

	Nombre de cas	p Valeur du Chi²	RC
SN possessif	15	0.26	0.43
Pronom personnel	310	0.0001	1.89
SN défini	206	0.0001	2.72
SN indéfini	61	0.0001	3.12
SN démonstratif	15	0.0001	4.24
Nom propre	212	0.0001	8.31

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 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 & 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.

3. 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 29. Il apparaît que ces expressions temporelles spécifiques associées à un nom propre obtiennent un RC très important (11.15).

Tableau 29 : RC des phrases débutant par un marqueur temporel du type « date (jour) » ou du type « le jour suivant », selon le sujet grammatical

	Nombre de cas	p Valeur du Chi²	RC
Pronom personnel	83	0.0001	3.52
SN défini	57	0.0001	4.18
SN indéfini	18	0.0001	4.44
Nom propre	79	0.0001	11.15

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.

4. Conclusions

Les analyses effectuées confirment une série d'hypothèses linguistiques au sujet de la fonction de marqueur de la segmentation des adverbiaux temporels. Tant les rapports des chances que l'indice issu de l'ASL mettent en évidence leur rôle de signal de rupture, 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

efficaces pour indiquer les ruptures temporelles que des expressions comme *En 1980*.

Cette recherche indique que les adverbiaux temporels sont de meilleurs signaux lorsqu'ils sont 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. Il serait 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 & al. (2006) sur la base d'un corpus journalistique. Ces hypothèses rejoignent les analyses de Vonk & 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.

Parmi les limitations de cette étude, il est impossible de ne pas mentionner la question du degré de généralisation des résultats. Le corpus utilisé est constitué 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. Celles-ci remplissent d'autres fonctions, plus complexes, de mise en évidence de la structure d'un texte (Charolles, 1997). L'étude des expressions temporelles nécessite une analyse linguistique plus fine, qui n'a pas été effectuée dans cette recherche.

En conclusion, ces résultats nous permettent d'envisager l'exploitation du cumul des marqueurs de rupture de thème pour les systèmes de segmentation automatique des textes. Ils autorisent 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.

Conclusions de la partie 2

Ces analyses de corpus nous apportent des résultats intéressants tant au niveau de la technique utilisée qu'au niveau des hypothèses linguistiques.

En ce qui concerne notre méthodologie, nous avons proposé deux indices pour identifier dans un corpus de textes des marqueurs de structure textuelle : le changement de paragraphe et un indice de cohésion lexicale issu de l'Analyse Sémantique Latente. Les analyses montrent que ces deux indices fonctionnent : les expressions récoltées dans la première analyse de corpus sont bien associées à des ruptures de thème. La deuxième analyse de corpus nous a permis de démontrer que l'indice de cohésion lexicale arrive à discriminer la capacité de marquage de différents types d'expressions temporelles. De plus, cet indice automatique donne des résultats semblables à ceux d'un juge humain ainsi qu'à d'autres indices automatiques. Il importe néanmoins de souligner les limites pratiques de cet indice basé sur l'analyse sémantique latente. 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 pour trois de nos études de corpus. 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 (comme c'est le cas dans la quatrième analyse de corpus), l'échantillon de phrases cibles devient de plus en plus petit, ce qui rend l'analyse de variance moins puissante. C'est une des raisons pour lesquelles nous avons toujours utilisé l'indice de cohésion lexicale couplé à l'indice de changement de paragraphe.

Dans sa version actuelle, l'indice basé sur la cohésion lexicale est moins satisfaisant que l'indice basé sur les paragraphes. Trois explications peuvent être proposées. 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 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 TextTilling (Hearst, 1997) ou CWM (Choi & 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 pas étonnant que ces deux décisions soient fréquemment prises simultanément.

Les analyses effectuées confirment également une série d'hypothèses linguistiques au sujet de la fonction de marqueur de la segmentation des différentes expressions étudiées. La fonction d'indicateur de la structure du discours des expressions référentielles et des adverbiaux temporels a été comparée dans la troisième analyse de corpus, ce qui nous a permis de les échelonner en fonction de leur capacité à marquer une continuité ou une rupture thématique. En ce qui concerne notre « fil rouge », l'anaphore possessive, la troisième analyse de corpus met en évidence que le SN possessif est considéré comme un indicateur de continuité thématique. Ce résultat demande une analyse plus fine. En effet, l'anaphore possessive peut être « doublement » anaphorique : le déterminant possessif renvoie à un possesseur déjà mentionné et le nom propre peut l'avoir été également. Dans notre corpus, nous avons détecté les formes « déterminant possessif + syntagme nominal » ; le principe de coopération veut que le possesseur soit donc connu, mais qu'en est-il du possédé ? Il se peut que ce soit la première mention du possédé sous cette forme. Il serait intéressant de

classer les expressions recueillies en séparant les expressions anaphoriques par le déterminant, des expressions doublement anaphoriques par le déterminant et le syntagme nominal.

Pour finir, dans la quatrième analyse de corpus, nous avons progressivement extrait les expressions temporelles les plus efficaces pour signaler une rupture thématique (telle que mesurée par nos indices). Ensuite, nous avons sélectionné les expressions référentielles qui apparaissent au début des phrases contenant ces expressions temporelles et mis en évidence 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, contrairement à l'hypothèse proposée par Vonk & al. (1992) qui soutient l'usage d'un seul type de marques de rupture à la fois. Le caractère additif de ce type d'expressions dans le marquage de la segmentation d'un texte confirme l'intérêt de développer des procédures d'identification de ruptures basées sur l'accumulation d'indices.