Vers une recherche automatique des marqueurs de la segmentation du discours

Sophie Piérard[1], Liesbeth Degand[2], Yves Bestgen[3]

[1]UCL/PSOR –1348 Louvain-la-Neuve – Belgique – sophie.pierard@psp.ucl.ac.be
[2]FNRS – UCL/LIGE –1348 Louvain-la-Neuve – Belgique – degand@lige.ucl.ac.be
[3]FNRS – UCL/PSOR –1348 Louvain-la-Neuve – Belgique – yves.bestgen@psp.ucl.ac.be

Abstract

To study linguistic expressions signalling thematic breaks in large text corpora automatic procedures for the identification of these breaks are indispensable. In the present study, we test the effectiveness of four indices of cohesion whose calculation can be automated. We show that these indices make it possible to differentiate between three categories of temporal segmentation markers.

Résumé

Afin de pouvoir étudier dans de grands corpus de textes le fonctionnement d'expressions linguistiques qui signalent les ruptures thématiques, il est indispensable de disposer de procédures automatiques capables d'identifier ces ruptures. Dans la présente étude, nous testons l'efficacité de quatre indices de cohésion dont le calcul peut être automatisé. Nous montrons que ces indices permettent de différencier trois catégories de marqueurs temporels de la segmentation.

Mots-clés : marqueur de la structure, segmentation automatique, anaphore, analyse sémantique latente.

1. Introduction[1]

L'objectif de cette étude est de proposer et de tester des procédures qui, à terme, devraient permettre la recherche automatique des expressions linguistiques qui signalent un changement de thème dans un texte. Ce genre d'expressions retient depuis de nombreuses années l'attention des chercheurs et une longue liste de candidats à cette fonction a été proposée : (van Dijk, 1982 ; Virtanen, 1992 ; Longacre, 1979). Les arguments empiriques présentés pour les soutenir sont généralement issus de l'analyse d'un nombre très réduit de textes (Stark, 1988 ; Virtanen, 1992). L'origine de cette limitation est d'ordre pratique. L'étude des marqueurs de la structure d'un texte présuppose la connaissance de cette structure. Classiquement, celle-ci est obtenue par une analyse linguistique fine ou par le recours à des juges auxquels on demande d'indiquer les ruptures thématiques qu'ils perçoivent (Bestgen, 1992 ; Passonneau et Litman, 1997). La complexité et le coût de ces procédures manuelles rendent l'étude de grands corpus impraticable. Pour dépasser cette limitation, il est nécessaire de s'appuyer sur des techniques qui permettent de déterminer automatiquement les changements de thèmes. De telles techniques ont été développées durant ces dix dernières années dans le champ de l'analyse automatique du langage. Elles visent à identifier les ruptures thématiques les plus importantes en s'appuyant principalement sur l'analyse de la cohésion lexicale entre les

[1] Cette recherche a bénéficié du soutien de la Communauté française de Belgique – Actions de recherche concertées et du Fonds de la Recherche fondamentale collective (FRFC/FNRS).

phrases successives d'un texte. Nous proposons d'employer ces techniques pour valider la fonction de marqueur de segmentation d'une expression linguistique en montrant qu'elle est fréquemment associée à des ruptures thématiques.

Pour qu'une telle approche soit viable, il est toutefois nécessaire de montrer que les techniques de segmentation sont suffisamment efficaces. Or, depuis quelques années, les procédures de validation sont devenues de moins en moins exigeantes. Alors que les premières études confrontaient la segmentation automatique aux jugements de lecteurs, on procède actuellement en soumettant à l'algorithme une série de textes artificiellement concaténés et en mesurant sa capacité à retrouver les frontières des différents textes (Choi, 2000 ; Utiyama et Isahara, 2001).

L'objet de la présente étude est d'évaluer la faisabilité de l'approche décrite ci-dessus. Il s'agit donc de répondre à deux questions : peut-on calculer des indices de continuité/discontinuité suffisamment fiables pour étudier le fonctionnement des marqueurs de la segmentation et, si oui, est-il possible d'automatiser leur calcul ? Comme situation de test, nous avons choisi un ensemble de marqueurs étudiés dans de petits corpus de textes. Leur intérêt majeur est qu'ils forment une échelle allant du signalement d'une rupture importante à celui d'une continuité importante (Section 2). Afin de montrer si la validité de cette échelle peut être confirmée par l'analyse d'un grand corpus de textes, nous avons comparé quatre indices de cohésion qu'il est possible de calculer, au moins d'une manière approximative, par des procédures automatiques (Section 3). La section 4 présente les résultats d'une expérience réalih 34 de.1009 Tw[(l on sont2 2pr hr4 d9(imatatre)3.9(-l)-48it

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 précèdent 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 ou encore

La seconde 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 adjectifs anaphoriques (personnel, possessif et démonstratif) en distinguant les anaphores intra-phrase ( ) des anaphores inter-phrases ( ) seules pertinentes ici. L'automatisation de cette procédure a simplement consisté en l'identification dans la phrase cible de l'ensemble des pronoms et adjectifs 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 mot peut endosser plusieurs fonctions grammaticales :

.

Les deux dernières mesures de continuité/discontinuité thématique ont été extraites par une procédure automatique basée sur l'Analyse Sémantique Latente (ASL : Landauer , 1998). Cette technique vise à construire un espace sémantique de très grande dimension à partir de l'analyse statistique des cooccurrences dans un corpus de textes. Cet espace sémantique peut ensuite être employé pour mesurer la proximité sémantique entre des phrases au moyen de la mesure classique du cosinus (Bestgen , 2003 ; Foltz , 1998). Récemment, l'analyse sémantique latente a été utilisée par Choi (2001) pour développer un algorithme de segmentation automatique plus efficace que plusieurs procédures classiques. Une présentation détaillée et une évaluation de l'efficacité de cet algorithme sont données par Bestgen (2004).

4. 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 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.

4.1. *Corpus et extraction des marqueurs en contexte*

Le corpus est composé de textes littéraires (romans, nouvelles et contes) extraits principalement des bases ABU et Frantext. Il contient 5 000 000 de mots. Les textes ont été découpés en phrases et lemmatisés au moyen du programme 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 :

– « A midi », « Vers midi », « A minuit », « Vers minuit », « A heure(s) », « Vers heure(s) » où 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 œuvres 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.

Pour l'estimation de la continuité/discontinuité par les chaînes référentielles, les deux indices suivants ont été calculés :
– 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 (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.

4.2. *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. 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 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.

	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 pour les trois conditions sont ordonnées comme attendu. Le connecteur 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 que de l'ancreur.

5. Conclusion

Même si cette étude présente plusieurs limitations (analyse d'un seul ensemble de marqueurs, caractères rudimentaires de certaines procédures d'estimation automatique des indices), ses implications nous semblent importantes à plus d'un titre. Tout d'abord, 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 . De plus, cette méthodologie peut être employée pour comparer d'autres marqueurs. Il serait, par exemple, intéressant d'opposer les ancreurs à des expressions comme « » qui ne semblent pas remplir la même fonction dans le discours (Bestgen et Vonk, 2000 ; Zwaan, 1996). Plus généralement, la méthodologie décrite devrait permettre l'utilisation d'une approche plus heuristique : plutôt que de vérifier si une expression linguistique est associée à des ruptures importantes, on pourrait rechercher les ruptures les plus importantes afin de déterminer les expressions linguistiques qui y occurrent fréquemment.

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.

Enfin, la présente étude s'inscrit dans le programme de recherches initié par Passoneau et Litman (1997). Ces auteurs ont proposé d'extraire la segmentation d'un texte sur la base de la cohésion lexicale, mais aussi de l'analyse des dispositifs linguistiques qui signalent la présence de changement de thème. Comme nous l'avons expliqué ci-dessus, nos indices référentiels et anaphoriques sont très proches de leur mesure des chaînes référentielles. Nos indices

présentent toutefois un avantage majeur : ils peuvent être, au moins partiellement, automatisés. Combiner en une seule procédure les indices référentiels, une analyse de la cohésion lexicale, par exemple par la procédure de Choi employée ici, et les marqueurs de la segmentation devrait permettre le développement d'un algorithme de segmentation particulièrement efficace.

Références

Bestgen Y. (2004). Analyse sémantique latente et segmentation automatique des textes. In

Bestgen Y. et Vonk W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. , vol. (42) : 74-87.

Bestgen Y., Degand L. et Spooren W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an explorative study. In Lagerwerf L., Spooren W. et Degand L. (Eds), Nodus Publikationen : 189-202

Bestgen Y. (1992). Le textomètre : un outil pour l'étude de la structure du discours et de son marquage. , vol. (32) : 141-167.

Choi F. (2000). Advances in domain independent linear text segmentation. In : 26-33.

Choi F., Wiemer-Hastings P. et Moore J. (2001). Latent Semantic Analysis for Text Segmentation. : 109-117.

Costermans J. (1998). . De Boeck Université.

Costermans J. et Bestgen Y. (1991). The role of temporal markers in the segmentation of narrative discourse. vol. (11) : 349-370.

Foltz P.W., Kintsch W. et Landauer T.K. (1998) The measurement of textual coherence with Latent Semantic Analysis. , vol. (25) : 285-307.

Kintsch W. et Van Dijk T.A. (1978). Toward a model of text comprehension and production. vol. (85/5) : 363-394.

Landauer T.K., Foltz P.W. et Laham D. (1998). An introduction to Latent Semantic Analysis. , vol. (25) : 259-284.

Longacre R.E. (1979). The paragraph as a grammatical unit. In Givón T. (Ed.), vol. (12), . Academic Press : 115-134.

Passonneau R.J. et Litman D.J. (1997). Discourse Segmentation by Human and Automated Means. , vol. (23) : 103-139.

Schmidt H. (1994). Version électronique disponible sur [http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html].

Segal E.M., Duchan J.F. et Scott P.J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. , vol. (14) : 27-54.

Stark H.A. (1988). What do paragraph markings do? , vol. (11) : 275-303.

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. In : 491-498.

van Dijk T.A. (1982). Episodes as units of discourse analysis. In Tannen D. (Ed.), . Georgetown University Press : 177-195

Virtanen T. (1992). . Åbo Akademi University Press.

Zwaan R.A. (1996). Processing narrative time shifts. , vol. (22) : 1196-1207.