À la pêche aux marqueurs linguistiques de la structure du discours

Sophie Piérard, Yves Bestgen

Université Catholique de Louvain – Belgique
Centre d'études du Texte et du Discours – PSOR

Abstract

This research aims at developing an automatic technique for finding expressions which signal the structure of a text. The technique consists (1) in a search for N-grams located at the beginning of sentences and (2) in the filtering of N-grams. For this second step, we propose to use two indices: the presence of a paragraph break and an index of lexical cohesion. This technique has been validated on two corpora: a training corpus and a test corpus. This experiment shows that the index based on the paragraphs is definitely more powerful than the index based on lexical cohesion.

Résumé

Cette recherche a pour objectif de développer une technique automatique de recherche d'expressions qui signalent la structure d'un texte. La technique employée consiste (1) en la recherche de N-grams situés en début de phrase et (2) le filtrage des N-grams. Pour cette seconde étape, nous proposons d'utiliser deux indices : la présence d'un changement de paragraphe et un indice de cohésion lexicale. Cette technique a été validée sur deux corpus : un corpus d'apprentissage et un corpus de test. Les résultats de l'expérience montrent que l'indice basé sur les paragraphes est nettement plus performant que l'indice basé sur la cohésion lexicale.

Mots-clés : structure de texte, marqueurs, N-gram, analyse sémantique latente.

1. Introduction

L'objectif de ce travail est de développer une technique automatique pour l'identification des expressions linguistiques qui signalent la structure de textes. Cette structure est déterminée par deux types de cohérence : la cohérence locale (entre les phrases, qui forment ainsi des segments) et la cohérence globale (entre les segments). Ce sont les éléments linguistiques qui signalent cette structure que nous étudions. Il s'agit tant d'éléments censés signaler les relations globales tels les adverbiaux cadratifs temporels ou spatiaux (« le matin», « en Belgique »), les connecteurs (« et ») ou les marqueurs métadiscursifs (« pour en revenir », « plus généralement ») (Charolles, 1997 ; Marcu, 2000 ; Segal *et al.*, 1991 ; Virtanen, 1992) que d'expressions censées signaler la continuité locale comme les marques de cohésion (les pronoms personnels « il », « elle », etc.).

Une meilleure connaissance de ces marqueurs est évidemment importante pour la linguistique, mais aussi pour la psycholinguistique et pour le traitement automatique des langues naturelles. En psycholinguistique, ces marques sont vues comme de véritables instructions, introduites par l'auteur, à destination du lecteur (ou de l'auditeur) afin qu'il construise la représentation mentale du texte la plus adéquate. Une meilleure connaissance de celles-ci devrait donc permettre d'améliorer l'intelligibilité de textes. Ces mêmes marques sont des sources d'information potentiellement très intéressantes pour les systèmes de

segmentation automatique des textes comme l'ont montré Passonneau et Litman (1997) et Beeferman *et al.* (1999).

2. Comment étudier ces marques ?

Pour étudier ces expressions, l'approche classique en linguistique, appelée sémasiologique ou inductive (Hansen, 1997), consiste à rechercher les occurrences d'une expression linguistique donnée afin de comparer sa fréquence dans différents corpus et d'essayer de déterminer les raisons de sa présence en certains lieux. Ce genre d'analyse peut être effectué de manière manuelle sur de petits corpus ou bien de la manière la plus automatisée possible sur de grands corpus. La limitation principale de cette approche est son caractère non heuristique, purement confirmatoire. Elle ne peut être employée que lorsqu'on dispose d'une liste d'expressions candidats marqueurs afin de vérifier si ceux-ci fonctionnent bien comme tels. C'est cette approche que nous avons employée dans une étude antérieure (Piérard, Degand et Bestgen, 2004) en montrant que diverses expressions adverbiales temporelles sont associées à des ruptures thématiques plus ou moins fortes.

Dans la présente étude, nous souhaitons tester une approche nettement plus exploratoire (par opposition à confirmatoire) et aussi plus heuristique : une analyse de corpus dirigée par la structure. Il s'agit d'essayer d'identifier automatiquement les expressions qui fonctionnent comme des marqueurs de la structure et, plus particulièrement, les expressions qui signalent la présence d'une discontinuité thématique entre des phrases. C'est ce que nous avons appelé la pêche aux marqueurs de la structure du discours. Dans la suite de ce rapport, nous décrivons la procédure que nous avons développée et une expérience menée sur deux larges corpus d'articles de presse qui en montre l'intérêt, mais aussi les limites.

3. 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.

3.1. Obtention des candidats marqueurs

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 (Degand et Bestgen, 2003 ; Hernandez et Grau, 2003 ; Schone et Jurafsky, 2001), mais aussi en statistique textuelle (Lebart et 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.

3.2. Filtrage des candidats marqueurs

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 et, de par leur statut de marqueurs de

discontinuité, ils apparaissent en des lieux de rupture thématique. La première contrainte nous a conduit à n'analyser que les N-grams qui se trouvent en début de phrases. La deuxième contrainte nous a conduit à développer deux indices de discontinuité thématique.

3.2.1. Le changement de paragraphe comme indice de discontinuité thématique

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). On peut donc penser qu'il aura tendance à introduire simultanément marqueurs de discontinuité et changement de paragraphe.

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 Chi² et le rapport des chances qui lui est associé (Howell, 1998 ; Piérard et Bestgen, 2005 ; Rogati et Yang, 2002).

3.2.2. La cohésion lexicale comme indice de discontinuité thématique

L'inconvénient majeur de ce premier indice est que les paragraphes remplissent d'autres fonctions discursives (Stark, 1988 ; Brown et 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, 2004, sous presse ; Choi *et al.*, 2001).

Ce second indice est issu de l'analyse sémantique latente, 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 (Landauer *et al.*, 1998). Cette technique permet d'inférer et de représenter le sens de mots sur la base de leur usage dans des textes afin de pouvoir estimer les similarités sémantiques entre des mots, des phrases ou des paragraphes. La similarité entre deux phrases est estimée au moyen de la métrique du cosinus. Plus deux phrases sont sémantiquement proches, plus leur cosinus est élevé.

Pour déterminer si un N-gram introduit une rupture thématique, on peut se baser sur les cosinus entre la phrase qui commence par ce N-gram, c'est-à-dire la phrase cible (p), et les deux phrases qui l'entourent : celle qui la précède (p-1) et celle qui la suit (p+1). Plus exactement, nous employons la différence entre ces cosinus. Si une phrase cible est en situation de discontinuité thématique, le cosinus entre cette phrase cible et celle qui la suit (cos[p,p+1]) devrait être plus grand que le cosinus entre cette même phrase cible et celle qui la précède (cos[p-1, p]) ; la différence entre ces deux cosinus (cos[p, p+1] - cos[p-1, p]) devrait donc être positive.

4. Expérience

4.1. Objectif

L'objectif de la présente étude est de comparer 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 un journal 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.

4.2. 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, ... 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, une étape indispensable pour identifier les N-grams apparaissant au début de celles-ci. 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. Les 300 premiers vecteurs ont été conservés et employés 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.

4.3. Résultats

4.3.1. 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[1]. On peut aussi s'attendre à ce que la différence des cosinus pour la première phrase d'un paragraphe soit plus grande que la différence des cosinus pour les phrases internes à un paragraphe. Ces prédictions sont heureusement vérifiées, puisque la

[1] 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.

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.

4.3.2. 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 *A 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 1, 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.

		Début de paragraphe		
		OUI	NON	Total
N-gram présent	OUI	14	4	18
	NON	102 766	346 932	451 998
	Total	102 780	346 936	452 516

Tableau 1 : table de contingence pour l'expression A l' étranger.

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.

4.3.3. Résultats

La figure 1 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.

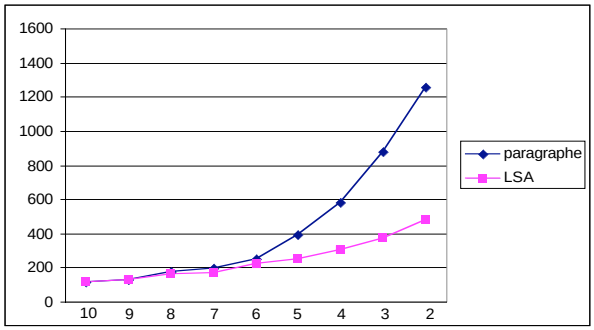

Figure 1 : nombre de N-grams sélectionnés selon le RC

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 2 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.

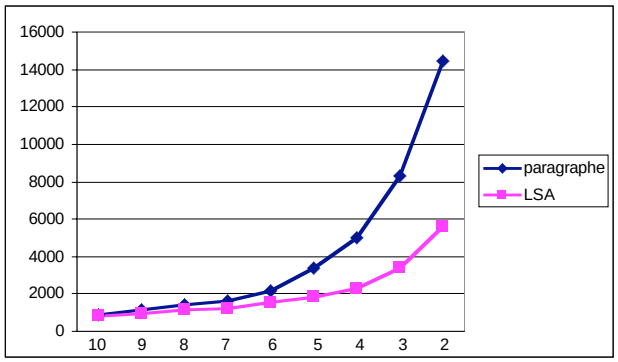

Figure 2 : nombre de phrases contenant les N-grams sélectionnés

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 3.

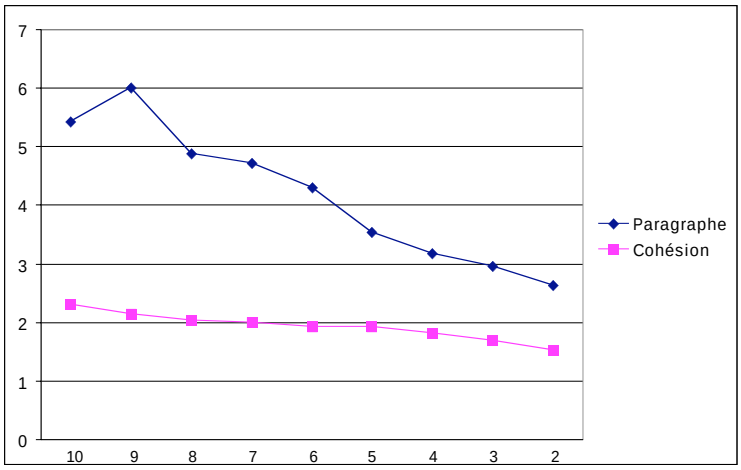

Figure 3 : rapports des chances moyens des expressions sélectionnées dans le corpus de test

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. Conclusion

Nous avons proposé deux indices pour identifier dans un corpus de textes des marqueurs de structure textuelle. Les résultats de l'expérience montrent que l'indice basé sur les paragraphes est nettement plus prometteur que l'indice basé sur la cohésion lexicale. Non seulement, il sélectionne un plus grand nombre d'expressions, mais en plus les expressions sélectionnées dans un premier corpus fonctionnent également comme marqueurs de rupture dans un second corpus.

Nous pensons que cette meilleure performance de l'indice basé sur les paragraphes s'explique par le fait qu'introduire un marqueur de la structure et un changement de paragraphe sont deux décisions de l'auteur d'un texte[2]. L'indice de cohésion lexicale peut très probablement être amélioré. Sa principale faiblesse, selon nous, réside dans son caractère local. Il indique seulement 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. Il serait donc intéressant de dériver des cosinus une mesure plus globale de la structure d'un texte par exemple en employant un algorithme de segmentation tel celui de Choi et al. (2001). Cela permettrait de

[2] Il est cependant possible qu'un relecteur ait modifié certaines décisions de l'auteur.

rechercher les ruptures thématiques les plus importantes et d'identifier les expressions qui les introduisent.

D'autres développements sont tout autant nécessaires pour que la pêche aux marqueurs de la structure se révèle fructueuse. Principalement, il serait utile de tenir compte, dès l'étape d'obtention des candidats marqueurs, du caractère peu figé de ces expressions. On pourrait par exemple regrouper les expressions voisines en se basant sur la présence d'éléments communs dans le même ordre et sur la proximité sémantique entre les éléments variables. Il serait aussi utile de développer des procédures de classification automatiques des expressions récoltées. À terme, l'objectif est d'inclure les marqueurs de discontinuité thématique dans un système de segmentation automatique des textes à la suite des travaux de Passonneau et Litman (1997) et de Beeferman et al. (1999).

Références

Beeferman D., Berger A. and Lafferty J. (1999). Statistical models for text segmentation. *Machine Learning*, 34 : 177–210.

Bestgen Y. (2004). Analyse sémantique latente et segmentation automatique des textes. In Purnelle G., Fairon C., and Dister A.(Éds.), *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, Presses universitaires de Louvain : 171-181.

Bestgen Y. (sous presse). Improving Text Segmentation Using Latent Semantic Analysis : A Reanalysis of Choi, Wiemer-Hastings and Moore (2001). *Computational Linguistics.*

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

Charolles M. (1997). L'encadrement du discours - univers, champs, domaines et espaces. *Cahier de recherche linguistique*, 6 : 1-73.

Choi F.,Wiemer-Hastings P. and Moore J. (2001). Latent semantic analysis for text segmentation. *Proceedings of NAACL'01* : 109–117.

Degand L. and Bestgen Y. (2003). Towards automatic retrieval of idioms in French Newspaper Corpora. *Literary and Linguistic Computing, 18* : 249-259.

Hansen M. M. (1997). Alors et donc in spoken French: A reanalysis. *Journal of Pragmatics, 28* : 153-187.

Hernandez N. and Grau B. (2003). Automatic extraction of meta-descriptors for text description. *International Conference on Recent Advances In Natural Language Processing (RANLP), Borovets, Bulgaria.*

Hofmann T.R. (1989). Paragraphs, & anaphora. *Journal of Pragmatics, 13* : 239-250.

Howell D.C. (1998). *Méthodes statistiques en sciences humaines*. Bruxelles, De Boeck Université.

Landauer T.K., Foltz P.W., and Laham D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, Vol. 25 : 259-284.

Lebart L. and Salem A. (1992). *Statistique textuelle*. Dunod.

Longacre R.E. (1979). The paragraph as a grammatical unit. In T. Givón, éd., *Syntax and Semantics, 12*, Discourse and Syntax, New York: Academic Press : 115-134.

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

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

Piérard S. and Bestgen Y. (2005). Identification automatique des marqueurs globaux du discours par l'analyse des expressions récurrentes. *Phraséologie 2005*, Louvain-la-Neuve.

Piérard S., Degand L. and Bestgen Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. In Purnelle G., Fairon C. and Dister A. (Éds.) *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve, Presses universitaires de Louvain : 859-864.

Rogati M. and Yang Y. (2002). High-Performing Feature Selection for Text Classification. *CIKM'02,* November 4–9, McLean, Virginia, USA.

Schmid H. (1994). Probabilistic Part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*.

Schone P. and Jurafsky D. (2001). Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? *Proceedings of Empirical Methods in Natural Language Processing,* Pittsburgh, PA.

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

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

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

Note des auteurs

Yves Bestgen est chercheur qualifié du Fonds national de la recherche scientifique (FNRS).

Cette recherche est financée par une "Action de Recherche concertée" du Gouvernement de la Communauté française de Belgique et par le projet FRFC n° 2.4535.02.