Comment évaluer les algorithmes de segmentation automatique ?

Essai de construction d'un matériel de référence.

Yves Bestgen, Sophie Piérard

Fonds national de la recherche scientifique, Université catholique de Louvain
{yves.bestgen ; sophie.pierard}@psp.ucl.ac.be

Résumé

L'objectif de cette recherche est d'évaluer l'efficacité d'algorithmes lors de l'identification des ruptures thématiques dans des textes. Pour ce faire, 32 articles de journaux ont été segmentés par des groupes de 15 juges. L'analyse de leurs réponses indique que chaque juge, pris individuellement, est peu fiable contrairement à l'indice global de segmentation, qui peut être dérivé des réponses de l'ensemble des juges. Si les deux algorithmes testés sont capables de retrouver le début des articles lorsque ceux-ci sont concaténés, ils échouent dans la détection des changements de thème perçus par la majorité des juges. Il faut toutefois noter que les juges, pris individuellement, sont eux-mêmes inefficaces dans l'identification des changements de thème. Dans la conclusion, nous évaluons différentes explications du faible niveau de performance observé.

Mots-clés : segmentation automatique, évaluation, accord interjuges.

Abstract

The objective of this research is to evaluate the efficacy of algorithms in identifying thematic breaks in texts. With this aim, 32 newspaper articles were segmented by groups of 15 judges. The analysis of their answers indicates that each judge, taken individually, is not very reliable, unlike the global index of segmentation derived from all the judges. If two algorithms tested are able to find the beginning of the articles when those are concatenated, they fail in the detection of the changes of topic perceived by the majority of the judges. It should however be noted that the judges, taken individually, are themselves largely ineffective in the identification of the changes of topic. In conclusion, we examine various explanations for the low level of performance observed.

Keywords: automatic segmentation, evaluation, interrater agreement.

1. Évaluer les algorithmes de segmentation automatique

La segmentation automatique de textes a pour objectif de localiser les changements de thème dans des documents. Ce type d'informations peut permettre l'amélioration de nombreuses applications en traitement automatique des langues naturelles comme l'extraction d'informations, le résumé automatique ou encore la navigation à l'intérieur de longs textes. Ces dernières années, de nombreux algorithmes ont été proposés (p. ex., Brants *et al.*, 2002 ; Choi, 2000 ; Ferret, 2002 ; Hearst, 1997 ; Ponte, Croft, 1997 ; Utiyama, Isahara, 2001) rendant encore plus important le double problème que pose leur évaluation. Il est en effet nécessaire de disposer non seulement d'indices efficaces pour mesurer les taux d'erreur, mais également d'une norme à laquelle la segmentation proposée par l'algorithme peut être comparée. En ce qui concerne la première question, l'indice proposé par Beeferman *et al.* (1999), et ultérieurement amélioré par Pevzner et Hearst (2002), semble faire actuellement l'objet d'un consensus.

La seconde question est nettement plus problématique. Comment déterminer les « véritables » changements de thème à identifier ? Deux approches sont principalement employées. La

première consiste à demander à des juges d'effectuer la même tâche que l'algorithme et donc à segmenter des textes de diverses origines (Hearst, 1997 ; Kozima, 1993 ; Passonneau et Litman, 1997). La seconde s'appuie sur un matériel artificiel obtenu en concaténant des textes, les changements de thème à identifier correspondant évidemment aux frontières entre ceux-ci. Depuis quelques années, cette seconde approche s'est imposée, en partie parce qu'un matériel de référence, conçu par Choi (2000), a été mis à la disposition des autres chercheurs pour évaluer l'efficacité de leur algorithme.

Évaluer un algorithme de segmentation au moyen de textes concaténés est une procédure parfaitement justifiée lorsque la fonction pour laquelle l'algorithme a été développé consiste à segmenter des séquences continues de brefs textes (Allan *et al.*, 1998 ; Ponte et Croft, 1997). Elle est, par contre, beaucoup plus discutable lorsque l'objectif est d'identifier les changements de thèmes à l'intérieur de textes. En effet, ceux-ci sont liés entre eux et au thème du texte en général. Prétendre qu'un algorithme efficace dans une situation le sera aussi dans l'autre est pour le moins imprudent comme le souligne Hearst (1997). Or, nombre d'applications de la segmentation automatique visent la mise au jour de la structure thématique d'un texte.

Ce constat nous a conduit à essayer de développer un matériel de test permettant d'évaluer l'efficacité de procédures de segmentation automatique dans l'identification des ruptures thématiques internes à un texte. Pour ce faire, nous avons demandé à des groupes de 15 juges d'identifier les changements de thème qu'ils percevaient dans 32 articles de différentes longueurs parus dans le journal *Le Monde* en 1995. La présentation de ce matériel de référence fait l'objet de la section suivante de ce rapport. Ce matériel est utilisé dans la troisième section pour comparer deux des principaux algorithmes de segmentation proposés dans la littérature : C99 (Choi, 2000) et TextTiling (Hearst, 1997). Si ces deux algorithmes sont capables de retrouver le début des articles lorsque ceux-ci sont concaténés, ils échouent dans la détection des changements de thème perçus par la majorité des juges à l'intérieur des textes. Il faut toutefois noter que les juges, pris individuellement, sont eux-mêmes inefficaces dans l'identification des changements de thème. Dans la conclusion, nous évaluons différentes explications du faible niveau de performance observé.

2. Constitution du matériel de référence

Pour construire le matériel de référence, nous nous sommes inspirés de la procédure employée par Hearst (1997). Les deux différences principales sont que les juges ont segmenté le matériel au niveau des phrases et non des paragraphes et que les articles à segmenter étaient issus d'un journal et non d'un magazine.

2.1. Sélection du matériel et obtention de la structure

Le matériel présenté aux juges a été sélectionné parmi les articles des trois derniers mois du journal *Le Monde* de 1995. Tous les articles d'au moins trois paragraphes et de minimum 200 mots ont été divisés en 4 groupes en fonction de leur longueur : entre 200 et 499 mots, entre 500 et 999, entre 1000 et 1499 et entre 1500 et 2500 mots. Huit textes ont été sélectionnés de manière aléatoire dans chacun de ces quatre groupes. Ces textes présentaient une grande variété de thèmes (politique, archéologique, philosophique, etc.). Ils étaient composés de paragraphes d'une longueur moyenne de 3.78 phrases (écart-type de 2.17).

Ces textes ont été mis en forme de manière à ce que chaque phrase soit séparée de la suivante par un retour à la ligne et qu'aucun indice typographique ne signale la présence d'un alinéa dans le texte original. Le titre de l'article a été conservé et placé dans un cadre afin de le

distinguer du texte à segmenter. Les sous-titres ont été supprimés. Huit carnets ont été créés, composés chacun de 4 textes (un texte par catégorie de longueur). Pour chaque carnet, deux ordres de présentation des textes ont été établis, l'un étant l'inverse de l'autre, avec la condition que le premier et le dernier texte de chaque carnet appartiennent aux deux catégories les plus longues.

Cent vingt participants (15 juges par carnet) ont pris part à cette recherche. Ils étaient tous étudiants en Bac 2 à l'Université catholique de Louvain et participaient à cette étude dans le cadre de travaux pratiques. Il leur était demandé de segmenter les textes en fonction des changements de thèmes qu'ils percevaient en traçant des lignes entre les phrases. Ils n'avaient pas d'indication quant au nombre minimal ou maximal de segmentations à effectuer.

2.2. Analyse des jugements

En moyenne, les juges ont segmenté toutes les 4.64 phrases, produisant donc des segments plus longs que les paragraphes. La question la plus importante à laquelle il est nécessaire de répondre porte sur le degré d'accord entre les juges. Cette question peut-être formulée de deux manières (Rosenthal, 1982). On peut s'intéresser à l'accord moyen entre deux juges. Est-ce que les juges, pris deux par deux, segmentent les textes aux mêmes points ? Il s'agit ici d'estimer la fiabilité (*reliability*) d'un juge. Si celle-ci est élevée, cela signifie qu'il n'était pas nécessaire d'interroger autant de juges différents puisque chacun d'entre eux apporte toute l'information. Cette fiabilité peut être estimée au moyen de différents coefficients tel le Kappa de Cohen calculé entre toutes les paires de juges possibles (Carletta, 1996).

La deuxième forme d'accord porte sur la fiabilité de l'indice de segmentation qui peut être dérivé des réponses de l'ensemble des juges. Cet indice est obtenu en comptant le nombre de juges qui ont segmenté entre chaque paire contiguë de phrases. Si nous avons interrogé autant de juges, c'est parce qu'on pouvait penser *a priori* qu'ils ne seraient pas systématiquement d'accord, que chaque juge percevrait la structure des textes imparfaitement, mais que les erreurs des uns seraient compensées par celles des autres et qu'au total, l'indice global refléterait adéquatement la segmentation des textes. C'est, au moins implicitement, la position prise par Hearst (1997) ou Passonneau et Litman (1997) lorsqu'ils ont décidé de ne prendre en compte que les segments marqués par plusieurs juges. Ce second indice d'accord inter-juges, qui mesure donc la fiabilité de l'ensemble des juges, peut être estimé au moyen du coefficient alpha de Cronbach (Nunnally, 1978).

Nous avons calculé ces deux types de fiabilité. La fiabilité d'un juge, estimée par le Kappa moyen, est de 0.42. Il indique que la concordance entre les juges est supérieure de 42 % à celles qui auraient été obtenues si les juges avaient segmenté d'une manière totalement aléatoire. Cette valeur est très faible puisqu'on considère classiquement qu'un kappa inférieur à 0.40 est insuffisant. La fiabilité de l'indice global, estimée par le coefficient alpha calculé pour chaque groupe de 15 juges, est en moyenne de 0.84. Cette valeur peut être interprétée de la manière suivante : si nous demandions à de nouveaux groupes de 15 juges, issus de la même population, de segmenter le matériel, les indices globaux dérivés de leurs réponses seraient corrélés en moyenne à 0.84 avec l'indice obtenu dans la présente étude. Un alpha égal ou supérieur à 0.70 indique une fiabilité satisfaisante (Nunnally, 1978).

En résumé, si les segments déterminés par un juge donné sont peu fiables, l'indice global de segmentation, qui est dérivé de l'ensemble des juges, est fiable. Correspondant à la proportion de juges qui ont segmenté entre chaque phrase, cet indice varie entre 0 et 1. Il est à noter qu'il est lié à la position des paragraphes dans les textes originaux. En moyenne, sa valeur est de 0.42 entre deux paragraphes alors qu'elle est seulement de 0.14 là où aucun paragraphe ne

débute. Ces deux valeurs sont statistiquement très différentes (test t pour comparaison de moyennes, $p<0.0001$). La fiabilité de cet indice global nous permettra donc de procéder comme Hearst (1997) ou Passonneau et Litman (1997) en considérant qu'un segment thématique a été identifié par les juges lorsqu'au moins 8 juges sur 15 l'ont marqué.

3. Évaluation de deux algorithmes de segmentation

3.1. Procédure

Le matériel de test a été employé pour évaluer l'efficacité de deux algorithmes de segmentation automatique basés sur la cohésion lexicale : C99 développé par Choi (2000) et TextTiling de Hearst (1997). C99 est fréquemment employé comme point de repère pour évaluer de nouveaux algorithmes sur la base de matériels composés de textes concaténés. TextTiling a par contre été évalué au moyen d'une procédure et d'un matériel comparable à celui présenté ci-dessus.

Si ces deux algorithmes présentent les trois étapes classiques des procédures de segmentation basées sur la cohésion lexicale, les implémentations respectives sont très différentes. Lors de la première étape, le document à segmenter est divisé en unités textuelles minimales. Il s'agit des phrases pour C99 et de segments de *w* mots pour TextTilling. La seconde étape consiste en l'estimation des similarités entre les unités minimales. Les deux algorithmes emploient le classique indice du cosinus entre des vecteurs de mots. C99 calcule cet indice pour toutes les paires de phrases, qu'elles soient ou non contiguës. TextTiling ne le calcule que pour les paires contiguës, mais emploie une procédure basée sur deux fenêtres mobiles qui s'étendent sur *k* segments de mots à gauche et à droite de l'espace inter-segment pour lequel la similarité est estimée. Enfin, la segmentation proprement dite est effectuée par C99 au moyen d'une procédure d'analyse en grappes (*clustering*) qui segmente répétitivement le document selon les frontières entre les unités minimales qui maximisent la similarité moyenne à l'intérieur des segments ainsi constitués. TextTiling, par contre, recherche les frontières pour lesquelles la similarité lexicale est faible comparée à celles des frontières environnantes.

Pour être déclarés cohérents par ces algorithmes, deux passages de textes doivent contenir des mots communs. Il s'agit d'une conception très restrictive de la cohésion lexicale. Afin de dépasser cette limitation, nombre d'auteurs ont proposé de prendre en compte des connaissances sémantiques complémentaires extraites de thesaurus ou de grands corpus de textes (par exemple, Choi *et al.*, 2001 ; Ferret, 2002 ; Kozima, 1993 ; Morris et Hirst, 1991). Nous avons donc également évalué une version de chaque algorithme prenant en compte ce genre de connaissances acquises par une analyse sémantique latente (ASL). Pour ce faire, un espace sémantique a été dérivé sur la base des articles parus dans le journal *Le Monde* en 1995, à l'exception des articles composant le matériel de test comme recommandé par Bestgen (1995, sous presse) et les trois cents premiers vecteurs propres ont été conservés.

Avant toute analyse, tant le matériel de test que le corpus pour l'ASL a été lemmatisé au moyen du programme TreeTagger de Schmid (1994) et un ensemble de mots fonctionnels ou très fréquents ont été supprimés. Lemmatiser le corpus à segmenter est une procédure classique en segmentation automatique de textes parce que cela permet d'accroître la proportion de mots identiques. Elle est particulièrement utile en français parce qu'elle permet de ramener au même lemme les nombreuses formes conjuguées d'un verbe. Il faut cependant noter que Bestgen (2004) a montré que lorsque la segmentation automatique s'appuyait sur un espace sémantique extrait par ASL, la lemmatisation du matériel n'apportait pas de bénéfice au niveau de l'efficacité de la segmentation. Elle permet néanmoins de réduire fortement la taille de la matrice termes x documents.

3.2. Résultats

Les deux algorithmes dans leur version de base et dans leur version ASL ont été soumis à trois tests. Le premier, le plus simple, consistait à identifier le début de chaque article lorsque ceux-ci sont placés les uns à la suite des autres dans un ordre aléatoire. Le second test leur imposait d'identifier les débuts de paragraphe à l'intérieur de chaque texte. Enfin, le dernier test consistait à identifier les segments marqués par la majorité des juges (au moins 8 juges sur 15). Les segments obtenus selon ce critère ont une longueur moyenne de 6.15 phrases.

Afin de simplifier les analyses, nous avons à chaque fois indiqué aux algorithmes le nombre de segments qu'ils devaient identifier. Les paramètres pour chaque algorithme ont été fixés en fonction des indications données par leurs auteurs (Choi *et al.*, 2001 ; Hearst, 1997).

L'efficacité des algorithmes a été évaluée au moyen de l'indice *WindowDiff*, développé par Pevzner et Hearst (2002), qui mesure la proportion de désaccords entre la segmentation de référence et celle produite par l'algorithme. Ces valeurs ont été comparées entre elles, ainsi qu'à deux niveaux de référence. Le premier est un niveau de base minimal : la performance atteinte par une procédure qui place en des points choisis aléatoirement le nombre attendu de ruptures dans chaque texte. Cette procédure a été effectuée 5000 fois pour chaque segmentation de référence. Le deuxième niveau de référence correspond à la performance moyenne des juges et donne donc une sorte de niveau maximal (Hearst, 1997). Il a été calculé en comparant pour la mesure *WindowDiff* les segments déterminés par chaque juge à ceux déterminés par les autres juges selon le critère indiqué ci-dessus.

Les résultats de ces analyses sont donnés dans le tableau 1. Les deux algorithmes et tout particulièrement dans leur version ASL sont efficaces pour détecter les débuts des articles. Par contre, leur efficacité dans les deux autres tests, l'identification des paragraphes originaux et celles des segments signalés par la majorité des juges, est à peine supérieure à celle obtenue avec un algorithme qui segmente d'une manière totalement aléatoire. Tout au plus observe-t-on une performance légèrement meilleure de TextTiling pour l'identification des paragraphes.

	Articles	Paragraphes	Segments
Niveau de base	0.53	0.52	0.51
Juges	n/a	n/a	0.49
C99	0.27	0.49	0.46
TextTiling	0.18	0.42	0.48
C99-ASL	0.10	0.46	0.47
TextTiling-ASL	0.12	0.41	0.49

Tableau 1. Taux d'erreurs (WindowDiff) des procédures de segmentation (n/a indique que ce test est non applicable)

Il faut noter que la différence la plus importante pour la segmentation en paragraphe, celle entre TextTilling-ASL (0.41) et C99 (0.49) n'est pas statistiquement significative pour un alpha de 0.01 (test t pour mesures répétées, les textes à segmenter formant les 32

observations). Dans le cas des segments, le taux d'erreurs des juges considérés individuellement est aussi médiocre que celui des algorithmes. Cette observation est à mettre en relation avec la faible fiabilité des évaluations de chaque juge. Notons enfin que ces analyses ont été effectuées également en différenciant les articles selon leur longueur sans que des différences liées à ce facteur apparaissent.

4. Discussion et conclusion

Force est de constater à la fin de cette étude qu'aucun des deux algorithmes évalués n'a été capable d'identifier les ruptures thématiques que des groupes de quinze juges ont mises en évidence. Comment expliquer ce résultat ?

On pourrait en premier lieu mettre en cause les juges. Les analyses ont montré qu'ils n'étaient que très modestement d'accord les uns avec les autres. Il se pourrait donc qu'ils n'aient pas effectué la tâche de segmentation avec toute la rigueur souhaitable. Trois arguments plaident contre une telle conclusion. Tout d'abord, contrairement à la fiabilité individuelle d'un juge, la fiabilité globale de l'ensemble des juges est très bonne (alpha moyen de 0.84). Ensuite, la localisation des segments est liée aux frontières des paragraphes qui sont supposés exprimer, au moins partiellement, la structure selon l'auteur de l'article. Enfin, s'il est exact que Hearst (1997) a obtenu des Kappa plus élevés, ceux-ci n'étaient en moyenne que de 0.65 alors même qu'elle avait demandé à ces juges une tâche probablement plus aisée : regrouper des paragraphes et non des phrases.

En deuxième lieu, les algorithmes pourraient être responsables du faible niveau de performance obtenu. Toutefois, mettre uniquement ceux-ci en cause semble peu crédible pour les deux raisons suivantes. D'une part, ces algorithmes sont parmi les plus efficaces développés en linguistique computationnelle pour réaliser ce genre de tâches. S'il est vrai que quelques études ont montré qu'ils pouvaient être surpassés, les gains sont souvent très faibles et ont été obtenus avec un matériel composé d'articles concaténés. D'autre part, il a été montré que les algorithmes ne sont pas moins efficaces que les juges pris un à un.

La troisième origine possible de cette faible performance est liée aux textes employés dans cette étude. Notons en effet que cette étude, contrairement à la très grande majorité des travaux dans ce domaine, a employé un matériel en français et que celui est issu, comme le souligne un relecteur, d'un des journaux les plus difficiles à comprendre. Cette troisième explication peut être formulée à deux niveaux de généralité. Tout d'abord, on pourrait incriminer spécifiquement les articles de journaux analysés dont la structure thématique est peut-être insuffisamment explicite et donc difficilement détectable tant pour les algorithmes que pour les juges. Seule une analyse linguistique fine de leur structure permettrait d'évaluer cette hypothèse pour autant qu'il existe une théorie de la structure d'un texte suffisamment précise pour une telle analyse. Notons toutefois que les articles ont été sélectionnés d'une manière totalement aléatoire et donc qu'une telle conclusion imposerait de définir avec une grande précision le type de textes que des algorithmes et que des juges peuvent segmenter efficacement. De plus, il reste à expliquer pourquoi globalement les juges sont fiables.

La version plus générale de cette troisième explication met en cause non le matériel spécifique de cette étude, mais la structure thématique des documents en général. Lorsque celle-ci est particulièrement explicite, par exemple lorsque le document est composé de textes concaténés, les algorithmes sont efficaces et un juge seul est assez fiable pour le segmenter. Lorsque la structure est peu explicite, les juges éprouvent des difficultés pour l'identifier d'une manière fiable et les algorithmes sont peu efficaces. Il s'agirait donc d'un continuum

sur lequel le matériel employé dans la présente étude se situerait du côté implicite. Dans une telle situation, il semble illusoire d'exiger des algorithmes de surpasser les juges.

Il reste à comprendre à quoi correspond l'indice global fiable qui peut être dérivé de l'ensemble des réponses des juges. Deux conceptions nous semblent possibles. Tout d'abord, on peut le voir comme le reflet de la structure hiérarchique d'un texte. En effet, comme aucune indication à propos du nombre de segments à délimiter n'était donnée aux juges, certains d'entre eux ont pu décider de n'isoler que les grandes sections alors que d'autres ont pu choisir de segmenter les textes en de nombreuses unités de petite taille. De telles différences réduisent fortement l'accord moyen entre deux juges. Elles nuisent nettement moins à la fiabilité de l'indice globale (Costermans et Bestgen, 1991). Cette première conception suggère une adaptation des mesures d'efficacité. En effet, si l'indice global traduit la structure hiérarchique, il est nécessaire de prendre cette dimension en compte lors de l'évaluation de l'adéquation d'une segmentation. Détecter une rupture hiérarchiquement plus importante devrait être davantage récompensé que détecter une rupture peu importante. Selon la seconde conception, le nombre de juges qui segmente en un point n'est pas lié à la structure hiérarchique du texte, mais au simple fait que certaines ruptures sont plus aisément localisables, par exemple parce qu'elles sont soulignées par une expression linguistique dont c'est la fonction comme *Par ailleurs* ou *En deuxième lieu*. Les juges seraient capables d'identifier fiablement ces ruptures, mais non les autres. Cette seconde conception remet en cause le recours à des juges pour obtenir la segmentation de référence de textes puisque ceux-ci ne sont fiables que pour une partie des ruptures. D'autres analyses et de nouvelles données sont nécessaires pour évaluer la pertinence de ces deux conceptions. Il serait, par exemple, intéressant de reconduire la tâche de jugement en étant plus explicite dans les consignes données aux juges. Il s'agirait par exemple de demander à certains juges de ne marquer que les ruptures les plus importantes ou obtenir des juges non seulement la position d'une rupture, mais aussi une évaluation de leur degré de certitude quant à la présence de cette rupture.

5. Remerciements

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

Références

ALLAN J., CARBONELL J., DODDINGTON G., YAMRON J. , YANG Y. (1998). « Topic Detection and Tracking Pilot Study. Final Report ». In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

BEEFERMAN D., BERGER A., LAFFERTY J. (1999). « Statistical models for text segmentation ». In *Machine Learning* 34 : 177–210.

BESTGEN Y. (2005). « Amélioration de la segmentation automatique des textes grâce aux connaissances acquises par l'analyse sémantique latente ». In *Actes de TALN 2005*. Dourdan : 203-212.

BESTGEN Y. (sous presse). « Improving text segmentation using Latent Semantic Analysis : A Reanalysis of Choi, Wiemer-Hastings and Moore (2001) ». In *Computational Linguistics* 32.

BRANTS T., CHEN F., TSOCHANTARIDIS I. (2002). « Topic-based document segmentation with probabilistic latent semantic analysis ». In *Proceedings of CIKM'02* : 211-218.

CARLETTA J. (1996). « Assessing agreement on classification tasks: The Kappa statistic ». In *Computational Linguistics* 22 : 249-254.

CHOI F. (2000). « Advances in domain independent linear text segmentation ». In *Proceedings of NAACL-00* : 26-33.

CHOI F., WIEMER-HASTINGS P., MOORE J. (2001). « Latent semantic analysis for text segmentation ». In *Proceedings of NAACL'01* : 109-117.

COSTERMANS J., BESTGEN Y. (1991). « The role of temporal markers in the segmentation of narrative discourse ». In *CPC / European Bulletin of Cognitive Psychology* 11 : 349-370.

FERRET O. (2002). « Using collocations for topic segmentation and link detection ». In *Proceedings of COLING 2002* : 260-266.

HEARST M. (1997). « TextTiling: Segmenting text into multi-paragraph subtopic passages ». In *Computational Linguistics* 23 : 33-64.

KOZIMA H. (1993). « Text segmentation based on similarity between words ». In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* : 286-288.

MORRIS J., HIRST G. (1991). « Lexical cohesion computed by thesaural relations as an indicator of the structure of text ». In *Computational Linguistics* 17 : 21-42.

NUNNALLY J. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill, New York.

PASSONNEAU R., LITMAN D. (1997). « Discourse segmentation by human and automated means ». In *Computational Linguistics* 23 : 103-139.

PEVZNER L., HEARST M. (2002). « A Critique and improvement of an evaluation metric for text segmentation ». In *Computational Linguistics,* 28 : 19-36.

PONTE J., CROFT W. (1997). « Text segmentation by topic ». In *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries* : 120-129.

ROSENTHAL R. (1982). « Conducting judgement studies ». In K.R. Scherer et P. Ekman (éds), *Handbook of methods in nonverbal behavior research.* Cambridge University Press, Cambridge : 287-361.

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

UTIYAMA M., ISAHARA H. (2001). « A Statistical model for domain-independent text segmentation ». In *Proceedings of ACL'2001* : 491-498.