

Transitions thématiques : Annotation d'un corpus journalistique et premières analyses

Alexandre Labadié¹ Patrice Enjalbert² Stéphane Ferrari²

(1) GETALP LIG, BP 53, 38041 Grenoble Cedex 9, France (2) Laboratoire GREYC, Université de Caen & CNRS, Bd Maréchal Juin, BP 5186 F-14032 Caen Cedex, France Alexandre.Labadie@imag.fr, Patrice.Enjalbert@unicaen.fr, Stephane.Ferrari@unicaen.fr

RÉSUMÉ

Le travail présenté dans cet article est centré sur la constitution d'un corpus de textes journalistiques annotés au niveau discursif d'un point de vue thématique. Le modèle d'annotation est une segmentation classique, à laquelle nous ajoutons un repérage de *zones de transition* entre unités thématiques. Nous faisons l'hypothèse que dans un texte bien construit, le scripteur fournit des indications aidant le lecteur à passer d'un sujet à un autre, l'identification de ces indices étant susceptible d'améliorer les procédures de segmentation automatique. Les annotations produites ont fait l'objet d'analyses quantitatives mettant en évidence un ensemble de propriétés des transitions entre thèmes.

ABSTRACT

Manual thematic annotation of a journalistic corpus : first observations and evaluation.

The work presented in this paper focuses on the creation of a corpus of journalistic texts annotated at discourse level, more precisely on a topic level. The annotation model is a classic segmentation one, to which we add *transition zones* between topical units. We assume that in a well-structured text, the author provides information helping the reader to move from one topic to another, where an identification of these clues is likely to improve automatic segmentation. The produced annotations have been subject of several quantitative analyses showing a set of linguistic properties of topical transitions.

MOTS-CLÉS : Structure du discours, segments thématiques, transitions thématiques, annotation.

KEYWORDS : Discourse structure, topical segments, topical transitions, annotation.

1 Introduction

On peut observer actuellement un intérêt croissant pour l'analyse de la structure du discours dans la communauté TALN, tant à des fins applicatives (indexation de documents, résumé automatique, ...), que pour des études de linguistique de corpus. Cet article est consacré à celles de ces approches qui tentent de saisir l'organisation du texte en termes de "blocs homogènes" successifs selon des critères thématiques (succession de thèmes tout au long du texte). Notre contribution concerne tant le modèle discursif lui-même (accent mis sur les zones de transition) que la constitution d'un corpus annoté en conséquence. Des ressources de ce type sont clairement nécessaires, comme "référence" pour l'évaluation de traitements automatiques ou pour des observations d'ordre linguistique. De fait, ce travail a été initié au sein du projet ANR Annodis, visant à l'annotation de textes selon plusieurs modèles discursifs (Péry-Woodley *et al.*, 2009)¹. Le

1. <http://w3.erss.univ-tlse2.fr/annodis/>

corpus est composé d'articles de journaux (articles de fond de politique et d'économie). La segmentation thématique est une tâche populaire en TALN (voir, par exemple (Morris et Hirst, 1991; Hearst, 1997; Choi, 2000)). Dans la plupart de ces travaux, les segments sont identifiés par leurs champs lexicaux et, malgré des résultats intéressants, on peut regretter que ces approches purement lexicales ne prennent pas en compte les signaux que les auteurs insèrent souvent dans le but d'informer le lecteur de changements de sujet. Dans notre travail, nous faisons l'hypothèse de l'existence de *zones de transition*, entre segments thématiques, ayant cette fonction, dans l'intention d'étudier leurs propriétés. Outre un intérêt théorique (mis en évidence par de nombreux travaux en linguistique) l'étude des indices de transition thématique est susceptible d'améliorer les performances d'outils de segmentation automatique². Afin de tester notre hypothèse, nous avons donc demandé aux annotateurs d'indiquer de telles zones. L'article est organisé comme suit. Nous présentons d'abord le corpus et la procédure d'annotation, de concert avec le modèle d'annotation thématique. Nous détaillons ensuite une mesure d'accord inter-annotateurs originale : aucune mesure réellement satisfaisante (permettant en particulier de comparer la segmentation de plusieurs annotateurs, et non le simple accord avec une annotation de référence) n'étant à ce jour disponible. Puis nous considérons la question des "signaux" indiquant les transitions entre segments thématiques.

2 Corpus et procédure d'annotation

2.1 Textes, outils et annotateurs

Le corpus en vue est composé de textes journalistiques du journal *Le Monde*, année 1994. Ce choix s'explique à la fois par les objectifs applicatifs (résumé, indexation, etc.) et par la qualité linguistique du journal. Nous avons choisi au hasard 30 textes (principalement en politique et en économie) de différentes tailles. Le corpus totalise 46 689 mots et a été réparti entre trois catégories : Petits : moins de 1000 mots (15 textes) ; Moyens : entre 1000 et 3000 mots (10 textes) et Grands : plus de 3000 mots (5 textes). Chaque texte a été annoté par 3 annotateurs au sein d'un groupe de 5 annotateurs "naïfs"³ en utilisant la plateforme Glozz⁴ (Widlöcher et Mathet, 2009). Les textes ont été présentés aux annotateurs avec un minimum de mise en page (titre, sous-titres et paragraphes).

2.2 Modèle d'annotation thématique

Il existe une grande variété de méthodes de segmentation thématique. La plupart admettent qu'un segment thématique est une unité de texte qui est thématiquement cohérente et thématiquement distincte de la précédente et de la suivante. Certaines se concentrent sur la recherche des frontières (Beeferman *et al.*, 1999), d'autres tentent de regrouper des portions de texte selon leur cohérence lexicale (Choi, 2000). Mais, que ce soit pour l'apprentissage ou l'évaluation, les expériences s'appliquent surtout à des corpus obtenus artificiellement à partir de l'agrégation de petits textes, l'hypothèse étant que la frontière entre deux textes peut être assimilée à une frontière thématique. Cette hypothèse est régulièrement critiquée (Bestgen et Pirard, 2006; Labadié et Prince, 2008) car elle ne prend pas en compte la question de la cohérence globale du texte.

Divers travaux s'intéressent néanmoins aux marques linguistiques de ruptures thématiques. On pourra citer des courants de recherche intéressés à la notion de segment cadratif, dans la lignée de (Charolles, 1997), ou à la typologie des expressions référentielles (Asher *et al.*, 2006). (Piérard et

2. (Couto J. *et al.*, 2004), par exemple, exploite divers types "d'introducteurs thématiques".

3. Des étudiants en master d'informatique et de sciences du langage n'ayant pas de connaissance *a priori* de la tâche et rémunérés pour l'occasion.

4. <http://www.glozz.org/>

Bestgen, 2006) étudie la manière dont ces deux types de marques peuvent se combiner avec des phénomènes de cohésion lexicale. Dans le domaine du TAL, (Couto J. *et al.*, 2004), exploite divers types "d'introduceurs thématiques" pour contribuer à la segmentation, tandis que (Widlöcher *et al.*, 2006) intègre divers indices linguistiques dans une procédure de fouille de texte.

C'est à cette recherche des marques de ruptures que notre travail se propose de contribuer, en s'appuyant sur la constitution d'un corpus annoté susceptible de les mettre en évidence. Notre modèle est basé sur l'hypothèse que, dans un texte bien construit, les ruptures thématiques brutales sont l'exception plutôt que la règle. Le plus souvent, des zones de transition aident le lecteur à passer d'un thème à l'autre. La zone de transition "parfaite", telle que nous la définissons, devrait comprendre à la fois une conclusion du thème précédent et une introduction du nouveau. Mais, même dans un texte correctement structuré, on peut trouver des transitions incomplètes, voir des frontières abruptes. La tâche d'annotation consiste alors à délimiter quatre types d'unités qui, par souci de simplification, doivent avoir la phrase comme unité atomique :

Les segments thématiques : Même si nous avons supposé que les transitions entre les thèmes sont, en principe, "floues" il convient, pour des raisons pratiques, de définir des frontières. Ces segments représentent les thèmes principaux et, le cas échéant, les sous-thèmes du texte. Par souci de simplification, nous nous sommes limités à un seul niveau de sous-thème.

Les introductions et conclusions thématiques : Elles correspondent à des portions de textes où l'annotateur repère des indices clairs indiquant que l'auteur introduit ou conclut un thème.

Les transitions "floues" : Si l'annotateur perçoit une transition, mais sans qu'une introduction ou une conclusion s'impose.

Nous n'avons volontairement pas donné aux annotateurs d'instructions sur ce que les "indices" de ces zones peuvent être, notre but étant précisément de les identifier.

3 Une mesure d'accord inter-annotateurs pour une tâche de segmentation

3.1 Contexte

Différentes mesures existent pour comparer des segmentations multiples portant sur un même texte. Ainsi, dans (Bestgen, 2009), l'auteur teste comparativement les efficacités de *WindowDiff* et de la distance de Hamming généralisée. Mais de telles mesures sont essentiellement adaptées pour comparer les résultats d'une segmentation automatique à une segmentation de référence. Elles ne peuvent être aisément généralisées à plus de 2 annotations concurrentes. Il existe par ailleurs la mesure classique Kappa (Fleiss, 1971), un standard qui permet de mesurer un taux d'accord à propos de la sémantique d'objets bien définis, mais ne prend pas en considération la notion de segment en tant que telle. Nous proposons de ce fait une nouvelle mesure, conçue pour répondre à notre problème de segmentation multi-annotateurs, directement applicable sur notre corpus.

3.2 La mesure d'accord

Notre mesure vise à proposer une évaluation adéquate des différences entre segments issus d'annotateurs distincts, ces segments ne constituant pas nécessairement un pavage complet du texte (cas des zones de transition par exemple).

Considérons dans un premier temps le cas de deux annotateurs, en supposant qu'une correspondance (alignement) ait déjà été effectuée entre leurs segments respectifs, que ce soit de manière automatique ou manuelle⁵. Lorsque nous comparons deux segments alignés, nous considérons

5. Le problème de l'alignement automatique de plusieurs segmentations est une question à part entière, qui dépasse le cadre de cet article.

les écarts-types (*StDev*) entre les bornes de début et les bornes de fin des deux segments. La somme de ces écart-types est ramenée proportionnellement à la moyenne des tailles des deux segments (*Avg*), par analogie avec le coefficient de variation (*StDev/Avg*) :

$$x = \frac{\text{StDev}(\text{pos}_{\text{init}}) + \text{StDev}(\text{pos}_{\text{end}})}{\text{Avg}(\text{size})}$$

. Ce calcul produit une valeur entre 0 et ∞ : 0 correspond à un alignement parfait (positions de début et de fin identiques) ; ∞ est une limite théorique correspondant à des positions divergentes (éloignement infini d'au moins l'une des bornes pour une taille de segment limitée).

Du fait de la difficulté d'interpréter un résultat entre 0 et ∞ , ce résultat est projeté sur un intervalle borné. Par analogie avec la mesure κ , nous avons choisi $] -1, 1]$. 0, le meilleur des cas, a pour image 1 et ∞ , le pire des cas, est projeté sur -1 . Nous avons accordé une attention particulière à la valeur $x = 1$. Cette valeur est atteinte lorsque la différence entre les positions est du même ordre de grandeur que la taille des segments elle-même, ce qui correspond, dans notre interprétation, à une limite entre accord et désaccord⁶, et nous conduit à projeter 1 sur la valeur "neutre" 0 de l'intervalle $] -1, 1]$. Ces contraintes sont satisfaites par l'utilisation pour la projection d'une fonction monotone, continue, construite à partir de la fonction tangente hyperbolique :

$$y = 1 - 2 \times \sqrt[2]{\frac{e^{\frac{x}{2}} - 1}{e^{\frac{x}{2}} + 1}}$$

. La mesure peut être facilement généralisée à n annotateurs en utilisant toujours les mesures statistiques écarts-types et moyennes entre n segments alignés (un par annotateur). Nous utilisons actuellement une moyenne arithmétique de ces résultats pour l'ensemble des annotations alignées à l'échelle d'un texte dans son intégralité.

3.3 Interprétation, alignement et usage

La projection sur $] -1, 1]$ avec par construction le 0 comme limite entre accord et désaccord permet une interprétation plus facile des résultats. Ainsi, toute valeur inférieure à 0 correspond à un désaccord qui sera plus important à mesure que la valeur se rapproche de -1 . À l'opposé, les valeurs supérieures à 0 représentent un accord qui ira en s'améliorant en avançant vers 1. Nous considérons que les valeurs supérieures à 0,3 sont satisfaisantes et deviennent très satisfaisantes au delà de 0,6. En effet, la fonction de normalisation possède une tangente infinie en 0, des écarts faibles font décroître rapidement les valeurs obtenues⁷.

Dans notre expérience, l'alignement a été effectué à la main de telle sorte que chaque segment de chaque annotateur soit aligné avec un segment de chacun des autres annotateurs⁸. Ainsi, si par exemple un annotateur "englobe" deux segments d'un autre annotateur, chacun de ces derniers sera aligné avec le premier. De plus, l'alignement a été fait "en aveugle", c'est-à-dire que la personne chargée d'aligner les segments n'a pris en compte que les position de début et de fin des segments, sans avoir connaissance du texte. Notre objectif n'est donc pas ici d'optimiser la mesure d'accord (comme (Mathet et Widlöcher, 2011)), mais de proposer un alignement continu de tous les annotateurs entre eux, même au détriment du score final.

6. Ce choix découle d'une observation préalable de différentes configurations de segments ; la normalisation de la mesure est donc construite pour reproduire cette interprétation de manière automatique.

7. Par exemple, si sur un segment annoté par dix annotateurs, un seul annotateur fait varier une seule des deux bornes du segment de 10% de la longueur du segment (les autres annotateurs étant en accord parfait), l'accord tombe à 0,83.

8. Seuls les scores d'accord des segments thématiques sont présentés, le faible nombre de zones de transition rendant l'interprétation d'un tel score sur ces dernières au mieux hasardeuse.

Le tableau 1 présente les scores obtenus en fonction de la catégorie de texte et pour l'ensemble du corpus. Pour chaque texte, ce score est la moyenne sur le texte des scores pour chaque segment (comme précisé en 3.2). Avec des scores allant de 0,16 à 0,7, mais majoritairement entre 0,3

| | Petits | Moyens | Grands | Global |
|----------------|--------|--------|--------|--------|
| <i>Minimum</i> | 0,36 | 0,18 | 0,16 | 0,16 |
| <i>Maximum</i> | 0,7 | 0,65 | 0,4 | 0,7 |
| <i>Moyenne</i> | 0,5 | 0,41 | 0,27 | 0,43 |
| σ | 0,11 | 0,12 | 0,1 | 0,13 |

TABLE 1 – Accord sur la segmentation du corpus

et 0,56, nous pouvons considérer, selon notre "grille de lecture", qu'il existe un accord global satisfaisant sur la structure thématique des textes. Un examen plus précis montre que si l'accord local parfait n'existe pas, la plupart des frontières proposées par les annotateurs ne sont pas éloignées de plus d'un paragraphe. En outre, il n'y a pas de cas de texte avec un désaccord moyen (< 0), ce qui laisse supposer que les lecteurs sont sensibles à des facteurs d'une nature relativement "objective". L'expérience menée sur les zones de transition, présentée plus loin, tend à confirmer ce point.

Une deuxième observation concerne la taille des textes : plus les textes sont petits, plus les scores sont élevés. Une raison évidente est la combinatoire des partitions possibles du texte, qui augmente avec la taille. Un facteur humain peut aussi intervenir, lié à la maintenance de l'attention de l'annotateur, qui peut être source d'erreurs pour les textes longs – au dire des annotateurs eux-mêmes. Enfin, une cause majeure de désaccord concerne la structuration en thèmes/sous-thèmes, i.e. la hiérarchisation des segments produits. Cette dernière conduit en effet certains annotateurs à diviser en plusieurs segments thématiques des portions de texte que d'autres considèrent comme un seul segment, ce qui fait, à dessein, considérablement chuter notre score d'accord.

4 Transitions thématiques

Notre objectif central concerne, rappelons-le, l'étude des transitions entre segments thématiques. Deux séries d'observations peuvent être faites à ce sujet, concernant d'une part leur perception par les annotateurs et d'autre part le repérage d'indices "de surface" de ces transitions.

4.1 Les transitions thématiques sont bien perçues par les annotateurs

Comme indiqué ci-dessus, les annotateurs ont dû délimiter les segments thématiques mais aussi des introductions, des conclusions et des transitions "floues". Une fois la phase d'annotation terminée, durant le processus de débriefing, nous avons demandé aux cinq annotateurs, entre autres choses, comment ils avaient procédé. La réponse a été claire et unanime : tout d'abord, chercher dans le texte où se situent les *changements de thèmes* ; ensuite, utiliser cette information pour identifier les segments thématiques. Cette façon de procéder leur semblait naturelle, et a confirmé nos hypothèses : les indications de changement de thème sont essentielles pour le lecteur, et clairement perçues par lui.

Les résultats chiffrés vont dans le même sens : 84% des segments thématiques ont une introduction, 39% une conclusion et 16% une zone de transition "floue". Sur l'ensemble des textes, presque tous les changements de thèmes sont donc accompagnés par une zone de transition, une introduction la plupart du temps.

4.2 Indices "de surface" pour les transitions thématiques

La question se pose alors de déterminer si ces zones possèdent des caractéristiques linguistiques spécifiques que le lecteur détecte, et quelles sont-elles. Outre l'intérêt théorique, l'identification de telles caractéristiques est susceptible de fournir des informations utiles pour un segmenteur automatique. Le problème est complexe, impliquant des mécanismes sémantiques et interprétatifs profonds, objet de divers travaux en linguistique comme rappelé dans l'introduction. Mais on peut aussi s'interroger sur la présence de formes *de surface* "réflétant" ces mécanismes. Afin d'étudier cette question nous avons fait les expériences suivantes.

Comme on pouvait s'y attendre, les introductions annotées coïncident presque toujours avec des changements de bloc, où "bloc" signifie "paragraphe" ou "sous-titre"⁹. Ainsi, en observant le corpus, nous avons considéré un certain nombre de traits distinctifs possibles et, pour chacun, nous avons établi une comparaison entre les introductions annotées et le début des paragraphes ne contenant pas d'introduction¹⁰. Trois hypothèses différentes ont été examinées dont les résultats sont présentés dans le tableau 2.

Les phrases introductives sont plus courtes. L'intuition est qu'elles peuvent opérer comme de véritables titres et transmettre l'information portée par le segment sous forme condensée. Par ailleurs il peut sembler judicieux à l'auteur (de manière consciente ou non) de signaler une nouvelle partie par une rupture de rythme dans le flot textuel. Les résultats montrent une réduction de 25% environ de la longueur des phrases des introductions thématiques par rapport à la première phrase des paragraphes classiques.

Les introductions thématiques aiment les segments détachés. Il est communément admis que les segments détachés en début de phrase ou, mieux, de paragraphe sont des positions thématiques privilégiées (Ho-Dac, 2008). Ces segments sont souvent séparés du reste du texte par une virgule, aussi avons-nous considéré une virgule en position pré-verbale comme une approximation acceptable pour identifier un détachement. Le ratio des détachements dans les introductions est de 0,4 contre 0,3 pour les autres paragraphes.

Les introductions thématiques ont des premiers mots préférés. La table 2 donne pour chaque catégorie morpho-syntaxique un ratio de ses occurrences en première position d'introduction thématique, par rapport aux autres paragraphes. On notera que la balise "adj" (adjectif) correspond en fait aux démonstratifs, "adv" (adverbe) généralement aux connecteurs du discours et "pun" (la ponctuation) presque toujours aux guillemets de citation. Les autres abréviations désignent respectivement les conjonctions, les déterminants, les noms, les prépositions, les pronoms et les verbes. Ces résultats suscitent plusieurs remarques : les adverbes (connecteurs de discours) sont sur-représentés, ce qui peut être interprété comme une influence de la structure rhétorique sur la structure thématique ; les conjonctions au contraire, et sans surprise, sont des marques de continuité, de même que les démonstratifs et les pronoms. Les déterminants et les noms sont sur-représentés, probablement en raison de leur fonction de focalisation sur un référent de discours (qui peut être nouveau en cas de changement thématique). Enfin, il n'y a dans notre corpus aucune introduction commençant par une citation.

Commentaires. Ces observations recourent partiellement certaines hypothèses formulées dans la littérature : par exemple les segments cadratifs vont apparaître dans notre première expérience en tant que constituants détachés et on retrouve l'idée commune selon laquelle les pronoms et les GN démonstratifs sont des marques de continuité. Ces éléments sont toutefois repris dans

9. Dans *Le Monde* un sous-titre n'est généralement pas le début d'une sous-partie, mais plutôt une sorte d'indicateur survenant *en son sein*, de sorte que les annotateurs ne les ont pas nécessairement étiquetés comme introductions.

10. 196 introductions pour 524 paragraphes.

un ensemble plus vaste de "candidats" (pour lesquels un examen linguistique attentif serait intéressant). Une différence est l'accent mis ici sur l'identification de marques de surface là où les études linguistiques proposent (avec pertinence, mais une exploitation informatique moins immédiate) des caractérisations sémantiques. Enfin notons que l'expérience semble plaider pour une certaine spécificité des paragraphes introducteurs de thèmes, ce qui peut questionner l'hypothèse formulée dans (Bestgen et Pirard, 2006) selon laquelle il serait possible d'utiliser de manière fiable le changement de paragraphe pour étudier les ruptures thématiques.

| | Introductions | Autres paragraphes | Différence en % |
|----------------------|---------------|--------------------|-----------------|
| Longueur des phrases | 17 | 24 | -23% |
| Détachement | 0,4 | 0,3 | +25% |
| Premier mot | | | |
| adj | 0,06 | 0,08 | -25% |
| adv | 0,09 | 0,05 | +45% |
| con | 0,01 | 0,04 | -75% |
| det | 0,4 | 0,33 | +20% |
| nom | 0,21 | 0,16 | +25% |
| pun | 0 | 0,04 | -100% |
| pre | 0,12 | 0,15 | -20% |
| pro | 0,9 | 0,11 | -18% |
| ver | 0,03 | 0,04 | -25% |

TABLE 2 – Indices de changements thématiques. Introductions contre les autres paragraphes.

5 Conclusion et travaux futurs

Dans cet article nous avons présenté un ensemble d'expériences relatives à la structuration discursive thématique, expériences appuyées sur la constitution d'un corpus annoté. De fait, le premier résultat positif est, selon nous, ce corpus lui-même, le manque de ressources de ce type étant généralement reconnu¹¹.

Nous avons également proposé une mesure d'accord multi-annotateurs adaptée à l'évaluation de l'accord entre segmentations multiples. Le défaut de telles mesures rend la constitution de tels corpus difficile et, bien qu'imparfaite (avec un l'alignement qui doit être fait *a priori* notamment), notre mesure à le double avantage d'être facile à mettre en œuvre et à interpréter.

L'accord inter-annotateurs semble suffisant (eu égard au caractère éminemment interprétatif du type de structure considéré) pour en faire un outil de recherche et les observations réalisées semblent confirmer la pertinence de la notion de zone de transition. Différents types d'indices de surface ont été mis en évidence, susceptibles d'aider au repérage de ces zones, donc à la segmentation automatique. L'intégration de ces indices dans une procédure de segmentation automatique à base lexicale pourrait également en constituer une validation "indirecte".

Une autre perspective consiste à étendre le corpus annoté afin de confirmer nos observations et, à terme, autoriser la mise en œuvre de procédures d'apprentissage.

11. Les annotations Glozz sont déportées, et peuvent être obtenues sur simple demande auprès des auteurs, charge au demandeur de posséder les droits sur le corpus du *Monde*

Références

- ASHER, N., DENNIS, P et REESE, B. (2006). Names and pops and discourse structure. *In Workshop on Constraints in Discourse*, pages 11–18, Maynooth.
- BEEFERMAN, D., BERGER, A. et LAFFERTY, J. (1999). Statistical models for text segmentation. *In Machine Learning*, volume 34, pages 177–210.
- BESTGEN, Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de textes ? *Actes de TALN'09*.
- BESTGEN, Y. et PIRARD, S. (2006). Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matériel de référence. *Actes de TALN'06*.
- CHAROLLES, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, 6:1–73.
- CHOI, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *Proceeding of NAACL-00*, pages 26–33.
- COUTO J., FERRET, O., GRAU, B., HERNANDEZ, N., JACKIEWICZ, A., MINEL, J. et PORHIEL, S. (2004). Régali, un système pour la visualisation sélective de documents. *Revue d'Intelligence Artificielle*, 18(4):481–514.
- FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.
- HO-DAC, L.-M. (2008). Exploring discourse organization through theme position. *19th European Systemic Functional Linguistics Conference and Workshop (ESFLCW) : Data and interpretation*.
- LABADIÉ, A. et PRINCE, V. (2008). Finding text boundaries and finding topic boundaries : two different task ? *Proceedings of GoTAL 2008*.
- MATHET, Y. et WIDLÖCHER, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *In TALN 2011*, volume 1, pages 247–258, Montpellier, France.
- MORRIS, J. et HIRST, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:20–48.
- PÉRY-WOODLEY, M.-P., ASHER, N., BENAMARA, F., BRAS, M., ENJALBERT, P., FABRE, C., FERRARI, S., HO-DAC, L.-M., LE DRAOULEC, A., MATHET, Y., MULLER, P., PRÉVOT, L., REBEYROLLE, J., VERGEZ-COURET, M., VIEU, L. et WIDLÖCHER, A. (2009). Annodis : une approche outillée de l'annotation de structures discursives en corpus. *Actes de TALN'09*.
- PIÉRARD, S. et BESTGEN, Y. (2006). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL*, 2(47):89–110.
- WIDLÖCHER, A., BILHAUT, F., HERNANDEZ, N., RIOULT, F., CHARNOIS, T., FERRARI, S. et ENJALBERT, P. (2006). Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte. *In DEFT'06*, Fribourg.
- WIDLÖCHER, A. et MATHET, Y. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *Actes de TALN'09*.