La sémantique d'un récit : état de l'art et perspectives

***Fionn Murtagh*[1,2], *Adam Ganz*[3]**
[1] Science Foundation Ireland, Wilton Place, Dublin 2, Irlande,
[2] Department of Computer Science, Royal Holloway, University of London Egham TW20 0EX, Angleterre
[3] Department of Media Arts, Royal Holloway, University of London Egham TW20 0EX, Angleterre
fmurtagh@acm.org

Résumé: On s'intéresse ici à la sémantique de l'information sous deux aspects : (i) l'ensemble de toutes les relations binaires, (ii) la reconnaissance et le suivi des changements ainsi que des anomalies. Dans le premier cas, on munit l'espace des textes ou des sous textes d'un côté, et l'espace des mots de l'autre côté, d' une métrique euclidienne dans un cadre commun. Dans le deuxième cas, on modélise l'information par une métrique ultramétrique. Une façon d'aboutir à une représentation euclidienne consiste à faire une analyse des correspondances, qui s'applique par exemple à des tableaux en entrée de fréquences. A partir de la représentation euclidienne, on munit l'espace de l'information d'une ultramétrique qui permet de construire une classification hiérarchique. En l'occurrence dans ce travail, on contraindra l'ultramétrique à respecter l'ordre induit par les séquences temporelles. Après une revue de l'existant pour l'analyse de la sémantique des scripts de films, on s'intéresse à la question suivante : serait-il possible d'analyser de façon analogue la sémantique de la littérature de la recherche.

Mots clés: Analyse des correspondances, classification ascendante hiérarchique, classification sous contrainte de contiguïté, analyse textuelle

Abstract We study two aspects of information semantics: (i) the collection of all relationships, (ii) tracking and spotting anomaly and change. The first is implemented by endowing all relevant information spaces with a Euclidean metric in a common projected space. The second is modeled by an induced ultrametric. A very general way to achieve a Euclidean embedding of different information spaces based on cross-tabulation counts (and from other input data formats) is provided by Correspondence Analysis. From there, the induced ultrametric that we are particularly interested in takes a sequential – e.g. temporal – ordering of the data into account. Following a review of approaches adopted in the analysis of filmscript we look at how similar approaches can be applied to the scholarly literature.

Keywords: Correspondence Analysis, hierarchical clustering, contiguity constrained clustering, text analysis

1 Analyse du récit

1.1 Introduction

L'analyse et la fouille des données doit relever principalement les défis suivants :

– Des grandes masses de données, sous forme textuelle et autre, doivent être exploitées comme base décisionnelle. L'analyse des correspondances fournit des réponses pour l'analyse de telles données multidimensionnelles, quantitatives ou nominales ; en un mot, hétérogènes.
– Les structures et les rapports évoluent dans le temps.

– Une toile complexe d'inter relations est à prendre en compte.
– Tous ces aspects doivent être considérés à partir des ensembles de données et de leurs flux.

Dans cette section nous prenons comme exemple le script du film, *Casablanca*. Dans la section 2, nous regardons plus en détail une scène de ce film, pour représenter et puis interpréter la sémantique du récit exprimé par cette scène. Dans la section 3 nous commençons à nous intéresser au problème suivant : dans quelle mesure peut-on utiliser notre méthodologie de l'analyse des récits pour une approche de l'analyse de la littérature de la recherche ?

1.2 Évolution dans la nature du cinéma et du drame

La production d'une émission d'une durée d'une heure dans une série télévisée coûte environ de 2 à 3 millions de dollars. La rédaction d'un scénario est typiquement soit une création autonome, soit une commande. Ensuite on réalise un épisode pilote qui sert de prototype.

Après avoir tiré ses origines dans un média, par exemple le cinéma, la télévision, un jeu, une réalisation on-line, ... on assiste au fait que de plus en plus souvent une série migre vers un autre support médiatique. On parle donc du multi-ciblage d'un scénario, "multiplay", ou de l'utilisation d'un scénario dans un cadre "à 360 degrés".

Les réalisations pour des plate-formes croisées aident à l'interactivité dans la dramaturgie. Un exemple important est ce qu'on appelle la télévision réalité qui combine beaucoup d'interactivité avec en correspondance peu de script.

De là découle l'intérêt d'une modélisation de la sémantique d'un script. On cherche à comprendre les structures sous jacentes, ses formes et ses niveaux d'expression. Avec l'évolution vers l'interactivité, on cherchera aussi à utiliser ce travail dans d'autres domaines où l'analyse à base de scénarios est importante. On pense par exemple à la planification stratégique dans le domaine des affaires, ou aux domaines de l'apprentissage et de la formation, ou bien encore à la politique de développement technologique et économique.

1.3 L'analyse des correspondances comme plate-forme d'analyse sémantique

Le point de départ pour l'analyse des correspondances [2, 9, 12, 14, 15, 18] est un tableau qui croise l'ensemble des scènes, en l'occurrence ordonné, par tous les mots utilisés dans le script.

Toutes les relations deux à deux des scènes et des mots sont représentées par leurs projections dans un espace euclidien, dit l'espace factoriel. La métrique euclidienne sert énormément la visualisation et à partir de là, l'interprétation.

Une nouveauté dans notre approche est l'importance accordée à l'orientation du récit. Comme on le verra ci-dessous (section 2) cela peut conduire à l'utilisation des corrélations plutôt que des projections.

Si la représentation factorielle tient compte de toutes les relations deux à deux, et de cette manière de la sémantique, il s'agit néanmoins d'une sémantique statique. Pour représenter les changements et les anomalies, nous utilisons une classification hiérarchique. Voyons comment cela fonctionne : dans la suite des agrégations, plus une scène est différentes des agrégations deux à deux effectuées précédemment dans la classification, plus l'indice et le niveau d'agrégation seront élevés. Nous avons donc besoin d'intégrer un élément nouveau dans notre approche. La classification doit tenir compte de l'ordre auquel la séquence des scènes est soumise. Nous disons que l'algorithme de la classification est sous contrainte de contiguïté, où en l'occurrence la contiguïté est fournie par la séquence. On trouvera plus de détails dans l'annexe et dans [11].

On voit que nous visons deux perspectives sur l'analyse sémantique : une vision globale des relations deux à deux entre les données, mais aussi l'expression du changement, par l'approche hiérarchique.

Pour l'étude du premier angle de vision sur la sémantique, un bilan des relations deux à deux est basée sur l'analyse factorielle. Les scènes sont munies d'une métrique euclidienne. On peut avec raison appeler cette perspective sur la sémantique *la géométrie de l'information* (terme utilisé par [20]). Avec la classification hiérarchique qui fait appel à une ultramétrique nous proposons sous cet aspect de dénommer la sémantique *la topologie de l'information*.

Sur le tableau de données en entrée, c'est la métrique du chi deux qui est appliquée. Puis dans l'espace factoriel produit par l'analyse des correspondances, on fait appel à la métrique euclidienne usuelle. Enfin, quand on arrive aux hiérarchies de classes, elles sont construites, elles, grâce à des ultramétriques.

Dans [13] nous allons plus loin pour développer des tests de type Monte Carlo pour la significativité des structures et des formes dans le script. Nous examinons également des césures et des changements dramatiques dans un script, à partir des classifications hiérarchiques tenant compte de la séquence des récits.

1.4 Première analyse du film Casablanca

La rédaction du script du film *Casablanca* n'était pas achevée quand le tournage a commencé en 1942. Il a été réalisé par Warner Brothers, entre mai et août de cette année-là. Le script était dû à J.J. Epstein, P.G. Epstein et H. Koch, à base d'un scénario qui date de 1940 (voir [4]).

Comme d'habitude dans un script, chaque scène contient du dialogue, associé aux noms des personnages. Il y a aussi des métadonnées dans chaque scène, soit concernant des consignes particulières, soit surtout avec les indications de Int (scène intérieure), Ext (extérieur), nuit, jour (ces éléments de métadonnées sont importants pour les informations concernant le déroulement temporel), Dans cette première analyse, nous avons construit un tableau croisant les 77 scènes successives par 12 attributs : Int, Ext, Day, Night, Rick, Ilsa, Renault, Strasser, Laszlo, Other (autres caractères mineurs), et deux lieux. Les lieux étaient : Rick's Café (sans distinguer entre "Main room", "Office", "Balcony", etc.) et tous les autres lieux, sans distinction.

Sur la figure 1 approximativement 34% (pour le premier facteur) + 15% (pour le deuxième facteur) = 49% de l'information, exprimée par l'inertie, se trouve expliquée. Des rapports entre les personnages (jouant un rôle de premier plan dans le film, bien sûr) peuvent être examinés sur cette représentation planaire. On voit également par exemple la proximité entre Rick's Café et Night et Int (ce qui se comprend facilement).

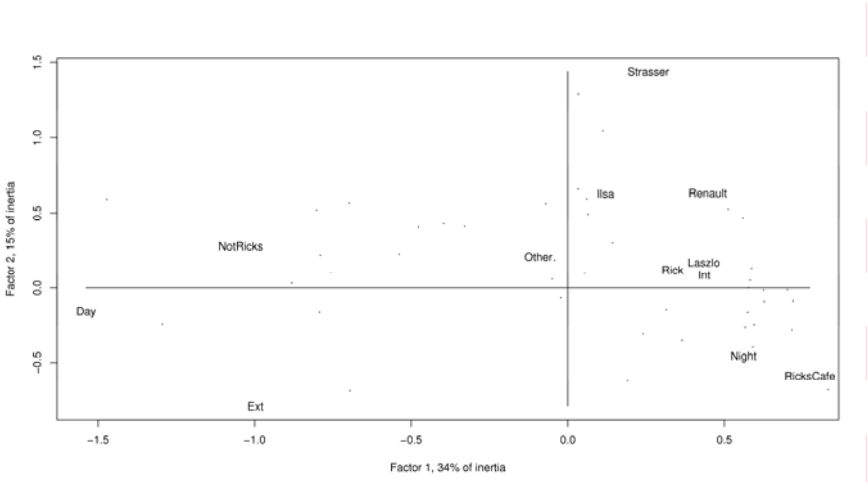

*Fig. 1: L'analyse des correspondances sur les 77 scènes du film **Casablanca** croisées par 12 attributs. Le tableau d'entrée contient des présences et absences. Les 77 scènes sont représentées ici par des points, pour faciliter la visualisation au premier abord.*

Sur la figure 2, où les scènes sont rangées par ordre séquentiel, on voit des changements atypiquement grands entre les scènes 9 et 10, et en allant de 29 à 30, et de 39 à 40 et puis à 41.

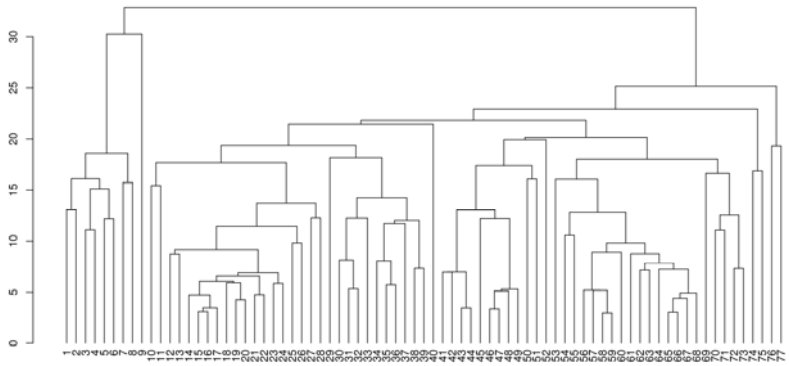

Fig. 2: La classification hiérarchique des 77 scènes. La séquence des scènes est respectée par l'algorithme ascendant hiérarchique. Le critère d'agrégation est celui dit du lien complet.

2 Analyse détaillée de la sémantique d'une scène de Casablanca

Nous allons maintenant procéder à l'analyse détaillée de la sémantique.

Pour McKee [10], *Casablanca* est "presque parfait" comme film. Attendu que la scène 43 est reprise en détail par McKee, nous allons la prendre pour l'étudier par l'analyse des correspondances. Pour reprendre McKee, cette scène est sous divisée en 11 "mesures" ou sous-scènes.

Ilsa et Rick se trouvent au marché en essayant de se procurer des visas de sortie. La séquence des sous-scènes est la suivante :

1. Dans la première sous-scène, Rick et Ilsa se trouvent au marché.
2. Dans les sous-scènes 2, 3 et 4, Ilsa le rejette.
3. Dans les sous-scènes 5 et 6, il y un rapprochement entre les deux.
4. Dans la septième sous-scène, scène 43, chacun se renvoie la culpabilité.
5. Il y a un saut imprévu dans le dialogue dans la huitième sous-scène : Ilsa annonce qu'elle va quitter bientôt Casablanca.
6. Dans la sous-scène 9, Rick dit que c'est 'une lâche' et Ilsa l'appelle 'idiot'.
7. Dans la sous-scène 10, Rick lui fait des avances.
8. Dans la sous-scène 11, l'apogée, Ilsa révèle qu'elle est mariée avec Laszlo. Rick est stupéfié.

Sur la figure 3 il y a une bonne représentation de l'évolution à travers la séquence des 11 sous-scènes. La sous-scène 8 se révèle comme très excentrique. Du côté positif des ordonnées apparaît l'éloignement entre Rick et Ilsa : voir les sous-scènes 2, 3, 4. En allant vers le côté négatif des ordonnées il y a rapprochement : voir les sous-scènes 4, 5, 6 et puis 9, 10. Dans l'espace initial on peut vérifier d'autres constatations énoncées par McKee. Lorsqu'on s'approche d'un moment culminant, les sous-scènes doivent être de longueur de plus en plus courte, à l'exception de la sous-scène finale exceptionnelle. Les longueurs des sous-scènes 7 jusqu'à 11 en terme de mots sont comme suit : 50, 44, 38, 30 et enfin 46. On trouve ici une bonne vérification de l'observation de McKee.

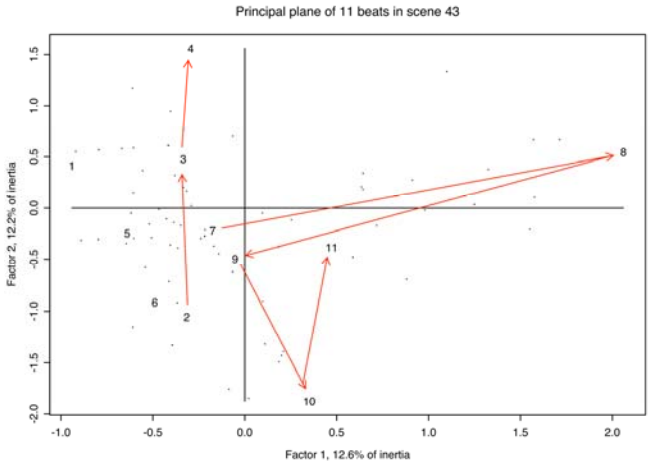

*Fig. 3: Le plan principal fourni par l'analyse des correspondances de la scène 43 de **Casablanca**. L'analyse porte sur 11 sous-scènes ou mesures ("beats"). Dans le texte nous discutons de l'évolution du dialogue à travers les sous-scènes 2, 3, 4 ; et ensuite à travers les sous-scènes 7, 8, 9, 10, 11.*

Une simulation de Monte Carlo peut être menée sur la scène 43. On commence avec la séquence donnée pour les 11 sous-scènes. Puis on considère 99 fois des séquences tirées au hasard. On utilise des permutations uniformément réparties. Cela permet de constater que dans la majorité des cas (83% ou plus des cas parmi les 99 cas aléatoires) la scène 43 avec sa séquence donnée est caractérisée par : une faible variabilité dans le mouvement d'une sous-scène à la prochaine, un tempo plus élevé dans la séquence donnée des sous-scènes, et un rythme moyen élevé.

La représentation planaire montrée sur la figure 3 explique approximativement 12.6% + 12.2% = 24.8% de l'inertie, et donc de l'information totale. Ce sont bien les changements en rapport à l'évolution du contenu qui constituent l'aspect le plus intéressant.

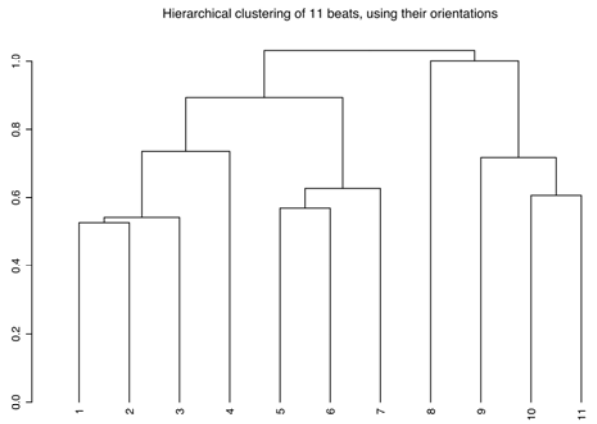

*Fig. 4: Une classification ascendante hiérarchique des sous-scènes de la scène 43 du film **Casablanca**. On utilise le critère d'agrégation du lien complet, sous contrainte de contiguïté relatif à la séquence des sous-scènes. Les données d'entrée sont les corrélations avec les facteurs, issus de l'analyse des correspondances, dans l'espace complet*

Sur la figure 4 on constate une césure en allant de la sous-scène 7 à 8, et de retour à 9. Il y a une césure moins caractérisée en allant de 4 à 5, mais néanmoins assez prononcée. Pour plus de discussion sur ces résultats, voir la référence [13].

3 Premiers pas dans l'application à la littérature de la recherche

Dans les dernières années l'évaluation de la recherche est devenue plus formelle et souvent basée sur les citations et des études bibliométriques. Le nombre de citations à un ouvrage est censé quantifier son importance et sa pertinence. Le fait que cette vue soit ou non correcte, ou bien même la mesure dans laquelle elle serait partiellement correcte, pourraient prêter à débat.

Les buts de la caractérisation bibliométrique comprennent les comparaisons de la performance internationale, la veille scientifique, et sous certains cas la répartition des ressources de financement (comme au Royaume-Uni dans le futur Cadre d'évaluation de la recherche, REF ou Research Evaluation Framework).

La tentative d'appliquer (au sens mathématique du terme) de manière automatique des scripts de films, sous forme textuelle, dans la prévision des recettes en salles a été suivi avec succès [6, 7, 8]. À partir d'un ensemble de caractéristiques qui permet d'obtenir ensuite un vecteur associé à chaque film, l'association des films au retour financier attendu est effectuée par un algorithme d'apprentissage.

Il nous apparaît donc que rien n'empêche, en principe, de traiter d'une façon analogue un article publié dans une revue de recherche. Il faudrait savoir quels sont les critères de réussite ou de retour. Rien n'empêche même de penser à une telle éventualité pour les propositions de financement de la recherche. Cela conduit à un outil pour aider un comité d'évaluation dans l'attribution du financement.

Comme point de départ avant d'aller dans cette voie de l'apprentissage automatique nous allons prendre en considération le contenu et la sémantique des articles publiés.

Dans ce contexte on pourrait se poser la question suivante: comment appréhender le *récit* fourni par un ensemble d'ouvrages (articles dans des revues de recherche) ? Dans quelle mesure pourrait-on analyser le *récit* fourni par une séquence d'ouvrages d'un seul auteur, et d'une équipe d'auteurs ou d'un laboratoire de recherche ?

Comme premier pas pour avancer dans cette voie nous allons examiner ici si une publication peut être représentée par les citations que l'on y fait.

3.1 Sélection et prétraitement des données

Dans un domaine particulier, celui du traitement des images neuronales sur la conscience visuelle ou des alternatives cognitives auprès d'aveugles dès les premières années, nous avons sélectionné 5 articles dans la revue *NeuroImage*. Ces articles sont les suivants : [19, 5, 1, 3, 17]. On voit qu'il y a des auteurs concernés par plusieurs publications [19, 5, 1] et un autre groupe d'auteurs [3, 17].

On fait d'abord quelques remarques sur le prétraitement de ces données textuelles. Les titres mineurs se retrouvent regroupés avec les paragraphes associés. Les titres majeurs sont ignorés. Les caractéristiques des articles sont résumées dans le tableau 1.

Article	No. de sections	N° de paragraphes	Mots totaux et	uniques
[19]	7	51	8067	1534
[5]	6	38	6776	1408
[1]	6	60	8247	1534
[3]	6	23	3891	999
[17]	7	24	5167	1255

Tab. 1: Propriétés des articles analysés

Parmi les différentes analyses présentées dans [14], nous allons nous focaliser ici sur le fait que les sommaires sont bien représentatifs des articles ; et, ce qui est plus intéressant vu le rôle joué en

bibliométrie, nous allons observer que les références aussi peuvent caractériser les articles correspondants.

3.2 Analyse des articles croisés par les mots, avec les sommaires en éléments supplémentaires

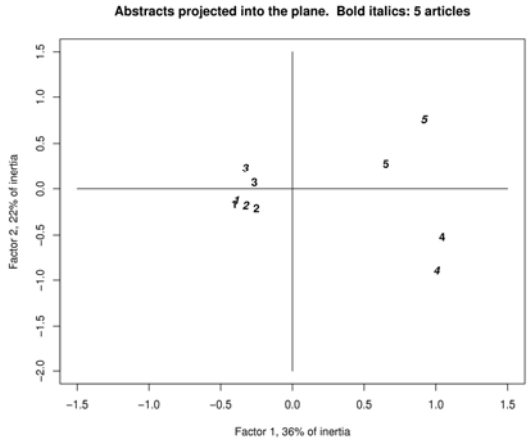

Fig. 5: Le plan principal de l'analyse des correspondances avec les sommaires en éléments supplémentaires. Les articles, qui sont les éléments principaux, sont imprimés en italiques.

Sur la figure 5, on voit que les associations entre les positions des sommaires et les articles sont assez bonnes, compte tenu dans les deux cas, de la grande différence de contenu.

3.3 Analyse des articles croisés par les mots, avec les références en éléments supplémentaires

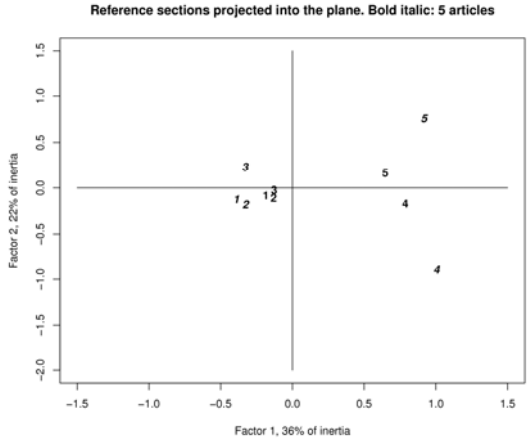

Fig. 6: Le plan principal de l'analyse des correspondances avec les références (les sections avec les références) en éléments supplémentaires. Les articles, qui sont les éléments principaux, sont imprimés en italiques.

En utilisant les sections avec les références, on a affaire à des noms d'auteurs, les titres des travaux cités, et les noms des journaux. Les numéros des volumes et des pages, ainsi que les dates, sont supprimés en raison de notre définition des mots à retenir. Il y a une différence fondamentale entre

l'article sans sa bibliographie et sa section bibliographique. Malgré cela la figure 6 montre que l'association des positions des éléments principaux et supplémentaires est cependant bonne.

4 Conclusion

Nous avons montré dans ce travail que l'enchaînement de l'analyse des correspondances avec une analyse complémentaire fournie par la classification hiérarchique est bien adapté pour faire ressortir les tendances essentielles de la matière d'un texte. Par cette approche, on peut donc faire ressortir la sémantique du texte et obtenir, de plus, une visualisation qui permet d'interpréter les tendances sous jacentes.

En appliquant cette méthodologie à la littérature, nous avons pris comme point de départ le contexte largement abordé de la bibliométrie. Nos résultats préliminaires indiquent que les références peuvent bien représenter le contenu des articles d'origine.

Références

[1] P. Arno, A.G. De Volder, A. Vanlierde, M.-Ch. Wanet-Defalque, E. Streel, A. Robert, S. Sanabria-Bohórquez and C. Veraart, *Occipital activation by pattern recognition in the early blind using auditory substitution for vision*, NeuroImage, 13, 632–645, 2001.

[2] J-P. Benzécri, *L'Analyse des Données, T I Taxinomie, T II Correspondances*, 2nd ed. Dunod, Paris, 1979.

[3] R.G. Bittar, M. Ptito, J. Faubert, S.O. Dumoulin and A. Ptito, *Activation of the remaining hemisphere following stimulation of the blind hemifield in hemispherectomized subjects*, NeuroImage, 10, 339–346, 1999.

[4] M. Burnett and J. Allison, *Everybody Comes to Rick's*, screenplay, 1940.

[5] A.G. De Volder, Hinako Toyama, Yuichi Kimura, Motohiro Kiyosawa, Hideki Nakano, A. Vanlierde, M.-C. Wanet-Defalque, Masahiro Mishina, Keiichi Oda, Kiichi Ishiwata and Michio Senda, *Auditory triggered mental imagery of shape involves visual association areas in early blind humans*, NeuroImage, 14, 129–139, 2001.

[6] J. Eliashberg, A. Elberse and M.A.A.M. Leenders, *The motion picture industry : critical issues in practice, current research, and new research directions*, Marketing Science, 25, 638–661, 2006.

[7] J. Eliashberg, S.K. Hui and Z.J. Zhang, *From storyline to box office : a new approach for green-lighting movie scripts*, Management Science, 53, 881–893, 2007.

[8] M. Gladwell, *The formula: what if you built a machine to predict hit movies?* , The New Yorker, 16 Oct. 2006. www.newyorker.com/archive/2006/10/16/061016fa_fact6

[9] L. Lebart, A. Salem and L. Berry, *Exploring Textual Data*, Kluwer, 1998. (L. Lebart and A. Salem, Analyse Statistique des Données Textuelles, Dunod, 1988.)

[10] R. McKee, *Story: Substance, Structure, Style, and the Principles of Screenwriting*, Methuen, 1999.

[11] F. Murtagh, *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg, 1985.

[12] F. Murtagh, *Correspondence Analysis and Data Coding with R and Java*, Chapman & Hall/CRC, 2005.

[13] F. Murtagh, A. Ganz and S. McKie, *The structure of narrative: the case of film scripts*, Pattern Recognition, 42, 302–312, 2009. (Discuté dans: Z. Merali, Here's looking at you, kid. Software promises to identify blockbuster scripts, Nature, 453, 708, 4 June 2008.)

[14] F. Murtagh, A. Ganz, S. McKie, J. Mothe and K. Englmeier, *Text sequence visualization using planar maps, hierarchical clustering and tag clouds : the case of filmscripts*, Information Visualization, sous presse, advance access online, 2009.

[15] F. Murtagh, Sixth Boole Lecture April 2008, *The Correspondence Analysis platform for uncovering deep structure in data and information*, Computer Journal, 2009, sous presse. advance access online, 2008 (doi :10.1093/comjnl/bxn045).

[16] F. Murtagh, *Semantics from narrative: state of the art and future prospects*, en preparation, 2010.

[17] M. Ptito, P. Johannsen, J. Faubert and A. Gjedde, *Activation of human extrageniculostriate pathways after damage to area V1*, NeuroImage, 9, 97–107, 1999.

[18] B. Le Roux and H. Rouanet, *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*, Kluwer, 2004.

[19] A. Vanlierde, A.G. De Volder, M.-C. Wanet-Defalque and C. Veraart, *Occipito-parietal cortex activation during visuo-spatial imagery in early blind humans*, NeuroImage, 19, 698–709, 2003.

[20] C.J. Van Rijsbergen, *The Geometry of Information Retrieval*, Cambridge University Press, 2004.