Un outil de coloriage de corpus pour la représentation de thèmes.

Pierre Beust[1]

[1]GREYC CNRS UMR 6072 & ModeSCoS – Université de Caen – 14032 Caen Cedex

Abstract

This paper presents a software called ThemeEditor. This tool provides a help to its users with the representation of their own semantic classes. This construction is realised through an interpretative analysis of an electronic texts corpora. A thematic coloring process plays the leading part in the software's principles. It consists in showing using several colors how the acquired (or under acquisition) semantic classes are set within the text.

Résumé

Cet article présente une application appelée ThemeEditor dont le but est de permettre une acquisition supervisée de classes sémantiques. Cette acquisition est réalisée dans le cours d'une tâche interactive d'analyse interprétative d'un corpus de textes électroniques. Le logiciel d'étude proposé met en œuvre un principe de coloriage thématique des documents du corpus. Il s'agit de mettre en évidence, en fonction des classes sémantiques acquises ou en cours d'acquisition, la répartition des thèmes et leurs différentes formes d'enchaînements.

Mots-clés : Segmentation thématique, Statistique textuelle, Traitement automatique du langage naturel, Logiciels pour l'analyse lexicale et textuelle.

1. Introduction

Notre recherche en traitement automatique des langues concerne la sémantique lexicale et la dimension thématique de la cohésion textuelle. Il s'agit dans cet article de présenter un logiciel interactif d'étude de corpus permettant à un utilisateur de mettre à profit cette cohésion pour construire les classes sémantiques qui l'intéressent. En cela, notre approche s'inscrit dans le courant issu des travaux en linguistique de Harris, Rastier ou encore Mel'cuk qui avancent que la construction de lexiques (à des fins de traitements automatiques ou d'analyses « manuelles ») est fondée sur une étude des usages des mots dans les productions langagières (textes, dialogues, …).

L'outil que nous proposons, appelé ThemeEditor, est basé sur une idée de coloriage que nous allons détailler dans une première partie de l'article. Dans une deuxième partie nous présenterons plus en détails l'application que nous avons développée. Enfin, nous expliquerons dans quels buts les classes sémantiques produites sont réutilisées par d'autres composants logiciels.

2. Le coloriage thématique.

De même que Pichon et Sébillot, 1999, nous entendrons ici par « thèmes » les sujets abordés dans un texte ou dans un corpus. Les thèmes seront représentés par des listes de mots indiquant le sujet en question. Le coloriage thématique est une façon d'identifier ces sujets dans les textes. Cette idée de coloriage provient directement d'une spécificité des langues

naturelles : introduire une très forte redondance dans le discours. C'est le cas au niveau morpho-syntaxiques avec, par exemple, les cas d'accords en nombre et en genre qui répètent plusieurs fois la même marque. C'est aussi le cas au niveau sémantique avec les récurrences de traits sémantiques (appelés sèmes) comme par exemple la répétition du trait /navigation à la voile/ dans l'énoncé :

Le skipper et son trimaran restaient encalaminés dans le pot-au-noir[1].

Depuis les travaux de Greimas, 1983, ces récurrences de traits sémantiques sont appelées isotopies. Le principe du coloriage thématique consiste à affecter une couleur à chaque isotopie et à « surligner » les mots du texte sur lesquels elles s'appuient. Le coloriage de textes électroniques permet ainsi de faire apparaître les différentes isotopies qui recouvrent un texte. On peut alors en examiner les répartitions au long du texte, leurs alternances et leurs enchaînements. De ce point de vue, le coloriage est aussi une méthode pour rendre objectif (et donc partageable) certains aspects fondamentaux des interprétations que l'on peut produire. En cela, l'outil que nous proposons s'inscrit dans le même courant d'étude que le logiciel PASTEL développé par Tanguy, 1997. A la différence de PASTEL conçu pour la visualisation des isotopies d'un texte, ThemeEditor est dédié à la construction de classes sémantiques à partir de corpus.

Plus qu'un constat remarquable sur la combinatoire des sèmes dans les chaînes linguistiques, l'isotopie est une base de l'interprétation des productions langagières (textes ou énoncés). Ainsi être capable de repérer les thématiques évoquées par un texte consiste à savoir y retrouver une ou plusieurs isotopies (Rastier, 1991, p. 11), c'est-à-dire affecter à certains mots du texte une même valeur thématique. Il s'agit alors de créer des signes linguistiques en faisant entrer certaines lexies du texte dans certaines classes sémantiques révélatrices de ce texte. En ce sens, il s'agit, dans une perspective de traitement automatique des langues, de fournir un outil pour l'acquisition de classes sémantiques qui ait en plus pour but d'apporter une aide à l'interprétation des textes électroniques.

> *En d'autres termes, l'interprétation ne s'appuie pas sur des signes déjà donnés, elle reconstitue les signes en identifiant leurs signifiants et en les associant à des signifiés. (Rastier et al., 1994)*

Compte tenu des phénomènes d'homonymie et surtout de la polysémie des langues naturelles qui touche massivement les domaines lexicaux courants, certains mots peuvent appartenir à plusieurs classes sémantiques révélant des domaines bien distincts. Ce serait par exemple, le cas du mot *avocat* que l'on pourrait aussi bien affecter à une classe sémantique indiquant le champ lexical des aliments qu'à une classe sémantique révélant un champ lexical juridique. Du point de vue de la méthode de coloriage, la question qu'il convient alors de se poser consiste à savoir quelle est la couleur à attribuer à un mot rencontré dans un texte si ce mot appartient à plusieurs classes. Dans un tel cas, l'heuristique considérée est celle qui tend à prolonger le plus possible les isotopies du texte (c'est-à-dire à favoriser la redondance). Ainsi, parmi ses couleurs possibles, on attribuera au mot la couleur la plus représentée dans le texte. Comme on le montre dans les deux exemples suivants, ceci constitue une méthode de désambiguïsation des mots polysémiques.

[1] Dans cet exemple (extrait de Rastier, 1991, p. 220), on a souligné les mots qui portent le trait sémantique /navigation à la voile/.

En faisant mon marché, j'ai vu des poireaux, des concombres et des avocats
En sortant du tribunal, j'ai vu mon avocat.

La méthode de construction des classes sémantiques par le coloriage thématique est essentiellement manuelle. C'est la différences avec des systèmes qui proposent automatiquement des thèmes par analyse des voisinages de mots (par exemple Pichon et Sébillot, 1999) ou bien qui proposent des ontologies issues de calculs de distances basés sur une analyse morpho-syntaxique (par exemple le système ASIUM de Faure, 2000). Comme pour PASTEL (Tanguy, 1997), notre méthode est basée sur une analyse interprétative de textes électroniques. Cette analyse est éclairée par des calculs statistiques réalisés par le logiciel. Partant d'un corpus de textes, le système permet de dresser des listes d'occurrences de mots à partir desquelles on enrichie les classes sémantiques qui en retour permettent le coloriage du corpus. L'utilisateur peut réitérer ce processus autant fois que nécessaire à la lumière des statistiques issues du coloriage (cf. Figure 1).

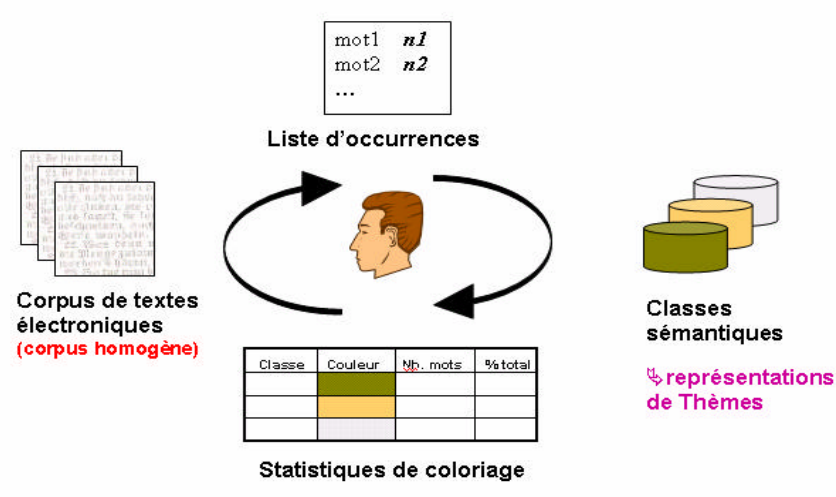


Figure 1 : Coloriage et acquisition de classes sémantiques

3. Un outil interactif pour l'analyse thématique de textes électroniques

L'outil ThemeEditor, est un logiciel d'étude disponible gratuitement en open source via Internet (cf. `http://users.info.unicaen.fr/~beust/ThemeEditor.html`). Il a été réalisé en Java par deux étudiants en maîtrise d'informatique à l'Université de Caen.

Il s'agit d'une application interactive qui permet à son utilisateur de créer et de modifier des thèmes, c'est-à-dire des classes de mots relevant d'un même domaine sémantique. Par exemple, la liste des 18 lexies (mots ou mots composés) suivante correspond à une définition succincte du thème Politique créé à l'aide de ThemeEditor :

élus, élu, Lionel Jospin, assemblée, assemblées, premier ministre, ministre, ministres, président, présidents, député, députés, référendum local, ministère, commune, communes, vote, gouvernement

Le principe du logiciel est de fonder la construction de ces thèmes sur des observations de productions langagières sémantiquement homogènes. Le matériau que nous avons utilisé pour expérimenter ThemeEditor est un corpus de textes électroniques représentant des dépêches d'agences de presses spécialisées dans les transactions financières.

L'interface de l'application présente 4 zones (cf. Figure 3). La première montre la liste des thèmes chargés. Si l'on ouvre l'un des thèmes on peut en consulter le contenu, c'est-à-dire la liste des mots retenus pour représenter le domaine sémantique visé. La deuxième zone de l'interface indique les noms des fichiers contenus dans le dossier qui contient le corpus d'étude. La troisième zone indique le résultat du calcul du nombre d'occurrences des différents mots du (ou des) document(s) sélectionné(s) dans la deuxième zone. La liste des mots est présentée par ordre décroissant de nombre d'occurrences. On peut ainsi extraire les mots les plus fréquents d'un texte, de plusieurs ou même de tous le corpus d'étude en fonction de la sélection de documents considérée (simple ou multiple).

Une interprétation de la loi de Zipf (Zipf, 1949) précise que si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquences décroissantes, on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste, ou autrement dit, que le produit de la fréquence de n'importe quel mot par son rang est constant : ce que traduit la formule f * r = C, où f est la fréquence et r le rang. Cette égalité, qui n'est vraie qu'en approximation, est indépendante des locuteurs, des types de textes et des langues. Selon ce principe appliqué à l'étude d'un corpus homogène (Giguet, 1998), on rencontrait approximativement dans les listes d'occurrences présentées par ThemeEditor tout d'abord les mots grammaticaux et par la suite les mots représentatifs du domaine de spécialité (Figure 2). On est donc en mesure de caractériser les principales classes terminologiques évoquées.

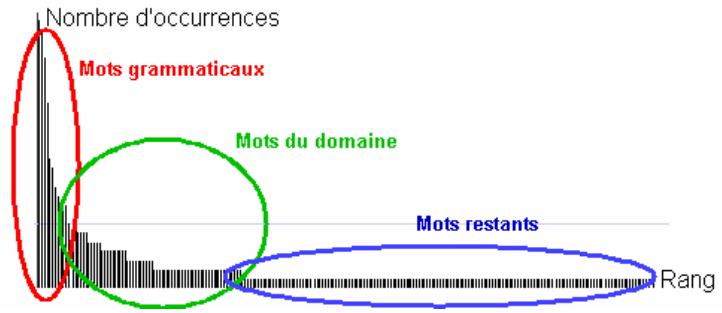
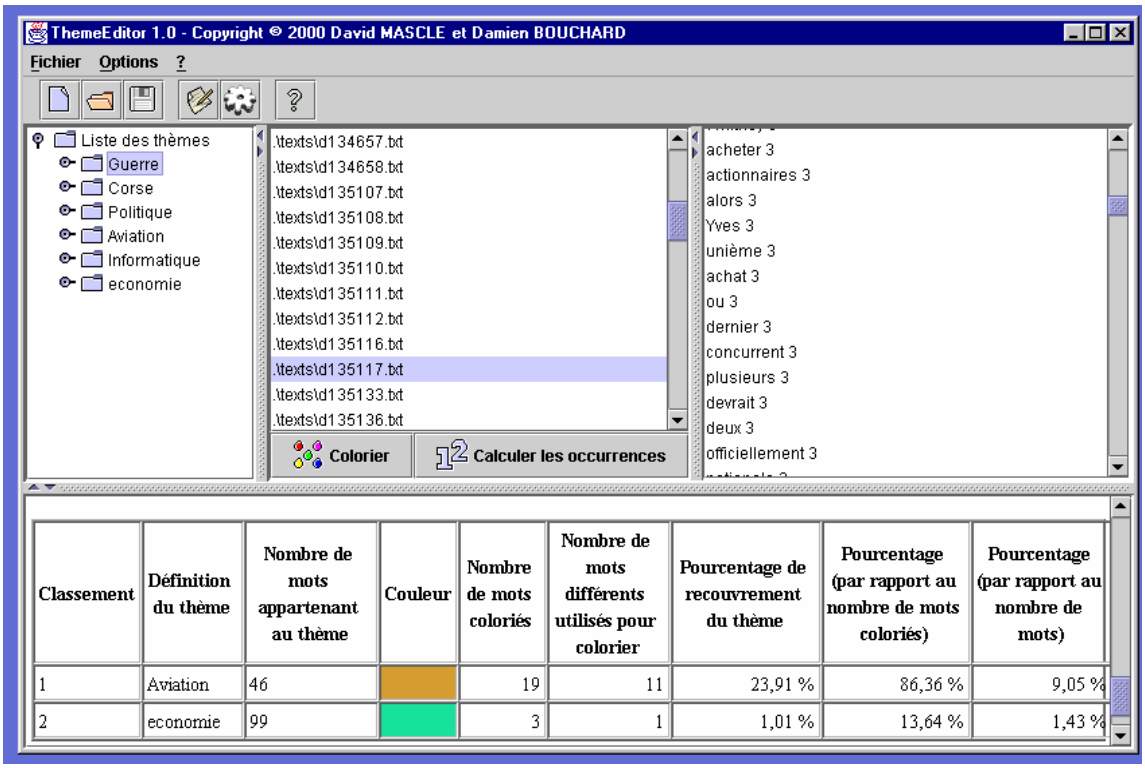


Figure 2. Loi de Zipf : Utilisation pour l'étude d'un corpus homogène

La quatrième zone de l'interface montre les résultats du coloriage sur le ou les textes sélectionnés. On trouve à la suite des textes où figurent colorés les mots des différents thèmes un jeu de calculs statistiques sur le coloriage réalisé, par exemple le classement des thèmes colorés dans la sélection avec pour chacun des thèmes le nombre de mots utilisés pour colorier ou encore le pourcentage que ces mots représentent par rapport à l'ensemble de la sélection.



Classement	Définition du thème	Nombre de mots appartenant au thème	Couleur	Nombre de mots coloriés	Nombre de mots différents utilisés pour colorier	Pourcentage de recouvrement du thème	Pourcentage (par rapport au nombre de mots coloriés)	Pourcentage (par rapport au nombre de mots)
1	Aviation	46		19	11	23,91 %	86,36 %	9,05 %
2	economie	99		3	1	1,01 %	13,64 %	1,43 %

Figure 3 : L'interface de ThemeEditor :
Résultats statistiques du coloriage du texte d135117.txt

Les classes sémantiques sont construites de manière incrémentale, soit en y entrant directement des lexies au clavier (cf. Figure 4), soit en cliquant sur les mots présentés dans la liste des occurrences pour les ajouter au thème préalablement sélectionné (cf. Figure 5). La possibilité de saisir directement au clavier des représentants de la classe terminologique est primordiale car des lexies composées ne peuvent apparaître dans la liste d'occurrences étant donné que l'interface Java que nous utilisons pour isoler les mots des textes utilise l'espace comme séparateur.

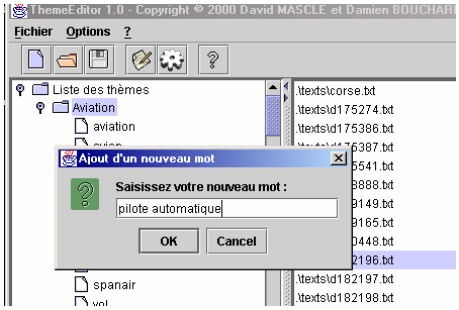


Figure 4. Saisie d'une lexie au clavier : ajout
de « pilote automatique » dans le thème aviation.

Une option de configuration de l'application permet d'utiliser une base de données lexicales (MulText[2]) via un serveur SQL pour ajouter au thème, en même temps que le mot sélectionné

[2] `http://www.lpl.univ-aix.fr/projects/multext/`

dans la liste, ses principales formes fléchies (par exemple les formes masculin ou féminin et singulier ou pluriel pour un adjectif ou encore l'ensemble des formes conjuguées pour un verbe à l'infinitif). Une autre option de configuration permet de définir un ensemble de mots qui ne seront pas pris en compte dans les calculs d'occurrences appelés mots vides. Ce sont des lexies qui ne relèvent pas de domaines sémantiques identifiables. Il s'agit principalement des mots grammaticaux qui apparaissent traditionnellement dans le haut de la liste des résultats de la loi de Zipf. Sur l'étude de notre corpus, on a pu remarquer que les mots vides forment approximativement 30% des mots des textes, ce qui représente à peu près le double du pourcentage de mots colorés (i.e. appartenant à un thème).

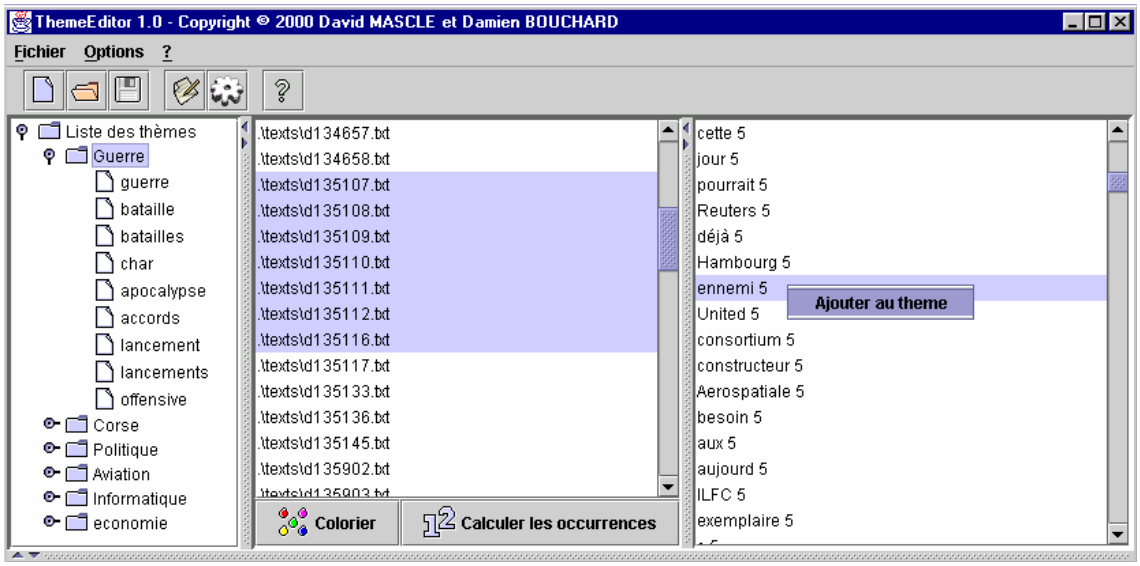


Figure 5 : La construction des thèmes : Ajout du mot ennemi au thème Guerre

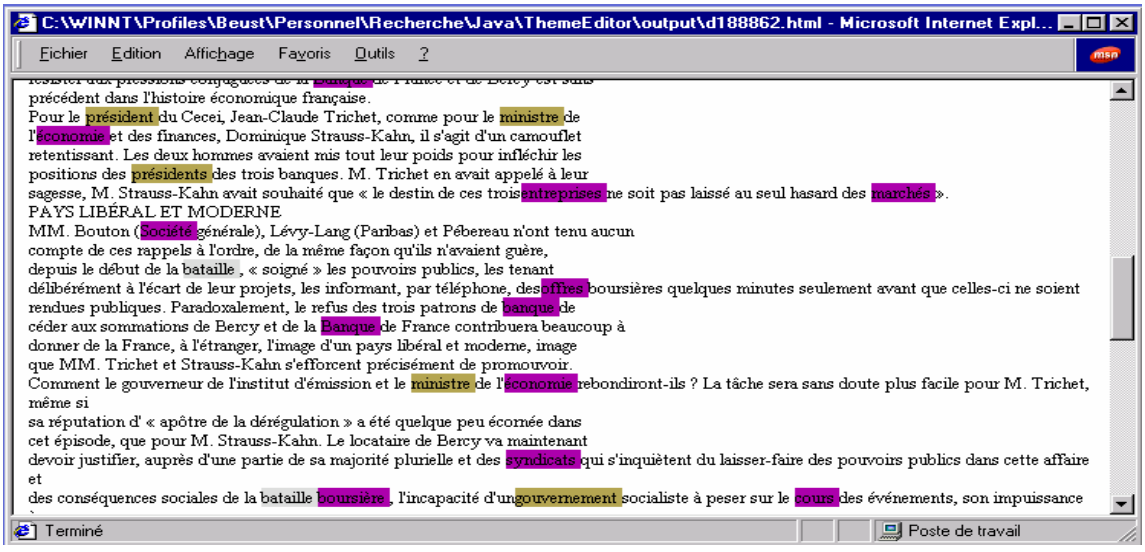
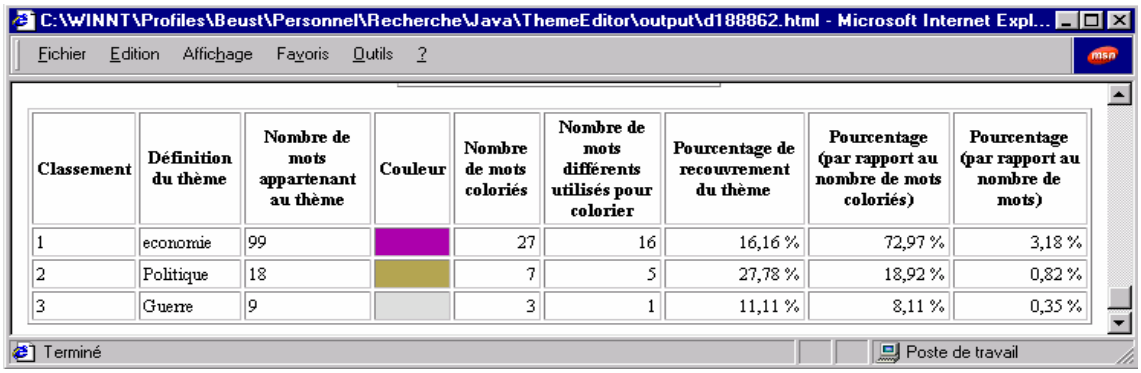


Figure 6 : Résultat d'un coloriage dans un fichier HTML :
Coloriage du texte d198862.txt

Les textes coloriés grâce aux classes sémantiques construites peuvent être visualisés dans la zone de l'application qui indique les statistiques mais ils peuvent aussi être enregistrés au format HTML et ainsi être visualisés dans un navigateur (cf. Figures 6 & 7). Ils peuvent également être enregistrés dans un document structuré par une DTD XML pour d'éventuels traitements automatiques ultérieurs. En plus du coloriage de fichiers textes, il est également possible de réaliser un coloriage en surchargeant la structure d'un document HTML, comme le montre la figure 8.



Classement	Définition du thème	Nombre de mots appartenant au thème	Couleur	Nombre de mots coloriés	Nombre de mots différents utilisés pour colorier	Pourcentage de recouvrement du thème	Pourcentage (par rapport au nombre de mots coloriés)	Pourcentage (par rapport au nombre de mots)
1	economie	99		27	16	16,16 %	72,97 %	3,18 %
2	Politique	18		7	5	27,78 %	18,92 %	0,82 %
3	Guerre	9		3	1	11,11 %	8,11 %	0,35 %

Figure 7 : Résultats statistiques du coloriage de la Figure 4:
Le coloriage du texte d198862.txt fait apparaître par ordre d'importance
les thèmes Economie, Politique et Guerre

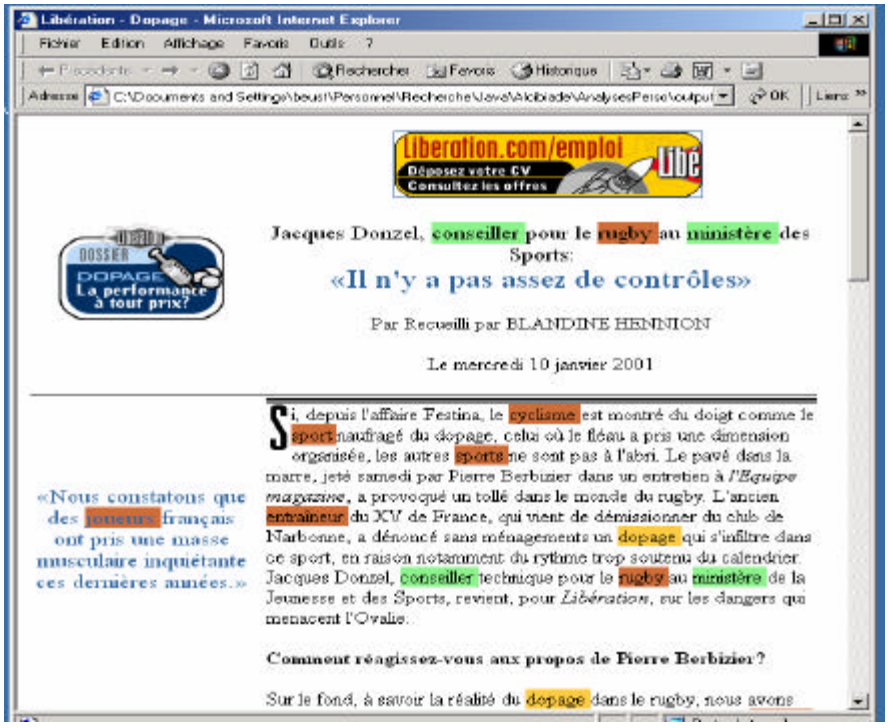


Figure 8 : Coloriage d'un document HTML

4. Utilisation des classes sémantiques produites

Nous avons fait une première expérimentation de l'outil que nous avons développé sur un petit corpus d'entraînement de 336 textes électroniques totalisant 260418 mots. Les principaux thèmes que nous avons construits dans cette expérimentation montrent en quoi il s'agit effectivement d'un corpus homogène. Ces thèmes sont l'aviation (133 lexies dans le thème) et l'économie (153 lexies dans le thème).

Les classes sémantiques produites sont enregistrées dans des documents électroniques répondant à une DTD XML spécifique. Ainsi, elles peuvent être réutilisées par d'autres traitements automatiques pour d'autres objectifs que le coloriage thématique.

4.1. Analyse statistique de flux documentaire

Une extension de ThemeEditor a été réalisée cette année sous forme de projet d'étudiants de maîtrise d'informatique. Elle concerne l'analyse de l'homogénéité thématique des flux documentaires. A la différence d'un corpus de textes, la spécificité d'un flux documentaire tient à son caractère dynamique dû à son ancrage temporel. En fonction de l'empan temporel considéré, le flux change par l'apparition et la disparition de documents. Cette dimension temporelle est à prendre en compte dans l'interprétation des documents du flux. Par exemple, on veut pouvoir rendre compte de l'apparition, des modifications et de la disparition de classes sémantiques en fonction de l'empan considéré. Dans ce but le logiciel développé présente différentes statistiques sur une période choisie : la densité du flux en terme de documents, l'évolution des thèmes dans la période(cf. Figure 9), la co-présence de certains thèmes dans cette période ou encore les proportions relatives des thèmes sur les documents de la période (cf. Figure 10).

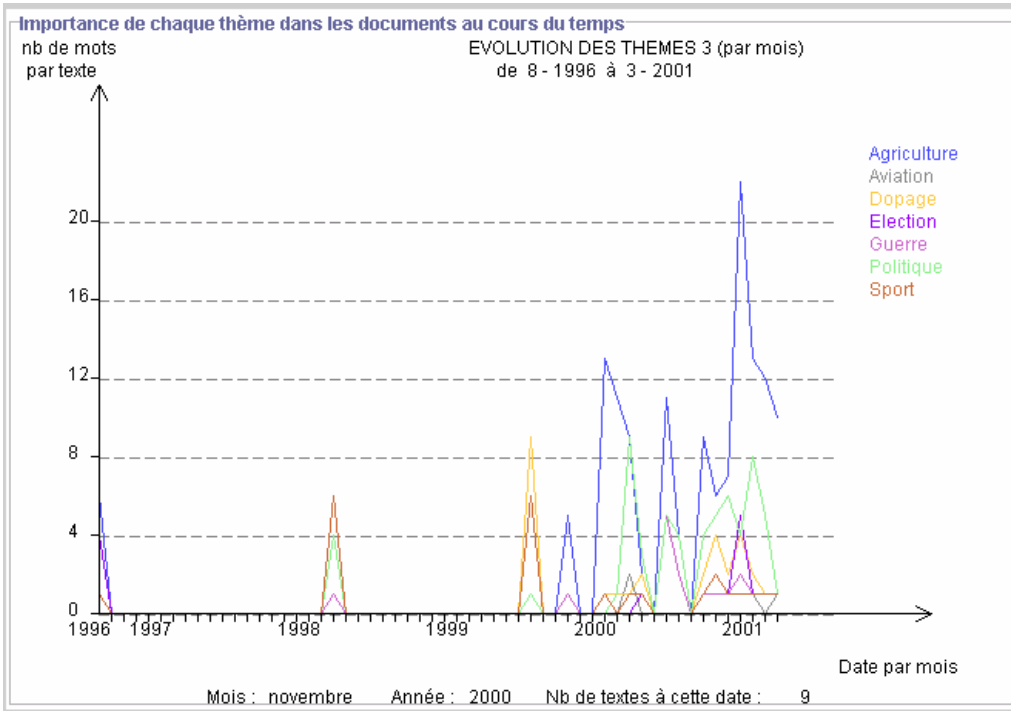


Figure 9 : Analyse de l'évolution des thèmes dans un flux documentaire

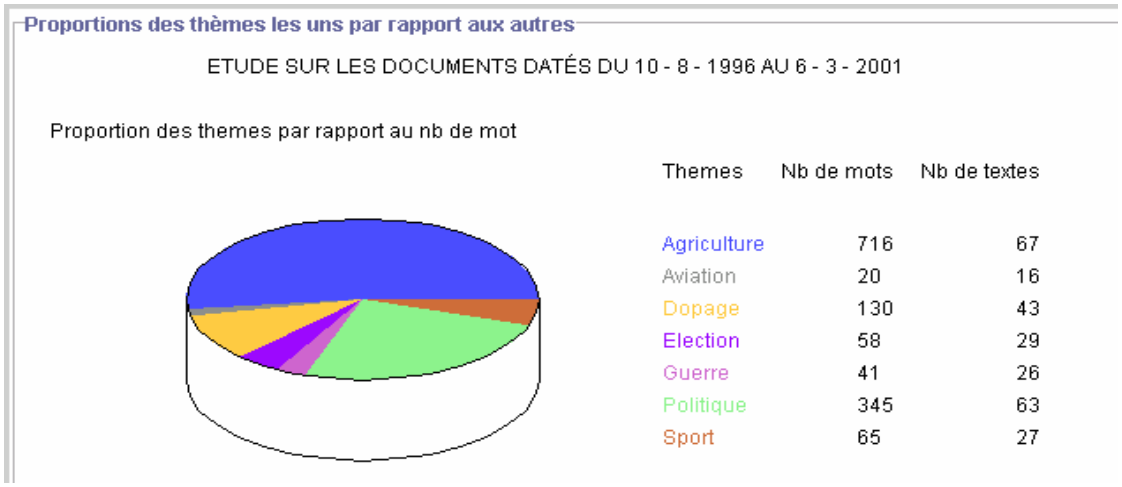


Figure 10 : Analyse de la proportion des thèmes dans un flux documentaire

4.2. Structuration des classes sémantiques

Les thèmes construits avec ThemeEditor peuvent fournir des descriptions en extension de champs terminologiques utilisés dans les domaines de spécialités évoqués par les textes analysés. Une utilisation possible de ces classes consiste à les étudier pour en extraire une représentation en compréhension et mettre ainsi en évidence leur structure lexicale.

Il est en effet un bon nombre de tâches où une représentation lexicale en terme de liste de mots n'est pas suffisante mais où les relations entre les termes (hyperonymie, hyponymie, synonymie ou encore antonymie par exemple) sont déterminantes. C'est le cas par exemple de la recherche documentaire sur Internet comme l'a montré Perlerin, 2000.

Pour construire des terminologies structurées à partir des thèmes et ainsi apporter une valeur ajoutée aux connaissances lexicales représentées pour un agent logiciel, nous utilisons l'application Anadia (Nicolle et al., 2001). Comme l'application ThemeEditor, Anadia est un logiciel d'étude réalisé en Java et disponible en open source via Internet (`http://users.info.unicaen.fr/~beust/anadia/`). Le logiciel met en œuvre un atelier interactif de catégorisation différentielle fondé sur la notion de valeur saussurienne (Saussure, 1915) dans lequel la structure d'une classe provient de la combinatoire des différences entre ses éléments. Les résultats produits sont des dispositifs qui rassemblent des tables indiquant une combinatoire de différences et des graphes appelés topiques montrant la proximité sémantique (deux sommets sont liés s'il ont une représentation identique à une seule différence près) entre les éléments d'une même classe. Ainsi dans l'exemple de la Figure 11, Anadia a permis de formuler un dispositif qui montre une structure différentielle possible pour un thème construit à l'aide de ThemeEditor. Ce thème évoque le domaine des matériels de l'informatique et comprend les lexies suivantes : *clavier, souris, appareil photo numérique, caméra vidéo, microphone, micro, écran, imprimante, écran avec haut-parleur intégré, écran avec haut-parleurs intégrés, haut-parleurs, HP, enceintes, enceinte, haut-parleur, support de données, clavier midi avec haut-parleur intégré, clavier midi avec haut-parleurs intégrés.*

	Type de medium	Mode d'utilisation
clavier, souris, appareil photo numérique	image et texte	entrée
caméra vidéo	image et texte et son	entrée
microphone, micro	son	entrée
écran, imprimante	image et texte	sortie
écran avec haut-parleur intégré, écran avec hauts-parleurs intégrés	image et texte et son	sortie
haut-parleurs, HP, enceintes, enceinte, haut-parleur	son	sortie
	image et texte	entrée / sortie
support de données	image et texte et son	entrée / sortie
clavier midi avec haut-parleur intégré, clavier midi avec hauts-parleurs intégrés	son	entrée / sortie

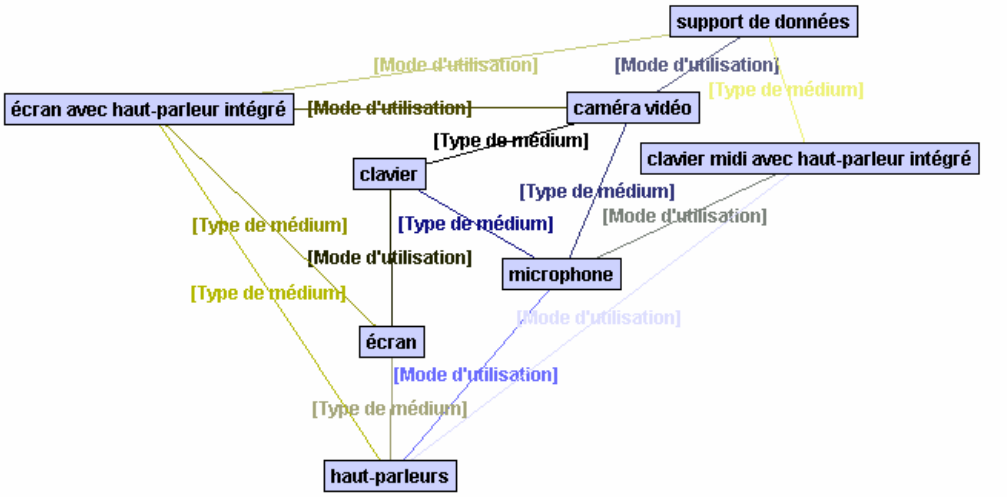


Figure 11 : Table et topique indiquant une structure différentielle pour la classe sémantique des matériels de l'informatique.

Comme les thèmes produits avec ThèmeEditor, les dispositifs Anadia représentant des terminologies structurées sont enregistrés dans des documents XML au format d'une DTD spécifique. Ils peuvent ainsi être utilisés à différentes fins. Nous en avons montré l'intérêt en ce qui concerne, l'analyse de la co-référenciation dans des séquences de dialogue (Beust, 2000), l'interprétation d'énoncés langagiers (Beust, 1998) pour un agent logiciel ou encore l'aide à la recherche documentaire (Perlerin, 2001).

5. Conclusion

A travers des outils tels que ThemeEditor et Anadia notre objectif est de mettre au point une plate-forme de logiciels d'étude pour l'analyse de la dimension thématique des corpus de documents électroniques.

Dans un soucis d'intégration de composants logiciels, il nous paraît primordial de développer ces outils avec des interfaces de communication de données sous forme de documents XML et de les diffuser en open source

La spécificité de ces différents outils réside en une approche anthropocentrée (telle que la définit Thlivitis, 1998), c'est-à-dire une approche dédiée à son utilisateur pour l'assister en fonction de ses propres besoins. Ainsi, ThemeEditor est un outil pour l'extraction rapide (une sorte de prototypage) de terminologies qui sont le reflet d'une analyse interprétative des textes étudiés par un utilisateur. Les représentations acquises par la machine sont principalement construites manuellement et de façon incrémentale d'une session de travail à une autre. En ce sens, elles ne représentent pas forcément un champ thématique de façon exhaustive et générale mais elles montrent en quoi, pour un utilisateur, certaines lexies d'un domaine ont une pertinence dans certains textes indiquant un ou plusieurs vocabulaires de spécialité.

Pour fournir une assistance logicielle dans des tâches complexes de veille technologique, il faudrait pouvoir élargir les terminologies produites pour ainsi pouvoir représenter la façon dont un thème est défini en langue plus que dans un corpus bien précis. Tout en conservant une approche anthropocentrée, on pourrait envisager de proposer à l'utilisateur de comparer les terminologies qu'il construit et structure manuellement avec les résultats de systèmes d'extraction automatique de terminologies comme ANA (Enguehard, 1993), ACABIT (Daille, 1995) ou encore le système décrit dans (Pichon et Sébillot, 1999), qui a par ailleurs l'avantage de se situer également dans le cadre théorique de la Sémantique Interprétative (Rastier, 1987). Une possibilité d'évaluation comparative serait de quantifier statistiquement le recouvrement et les différences entre les classes issues d'analyses manuelles et les classes issues de traitements automatiques. Ainsi, on pourrait essayer d'évaluer la valeur ajoutée de telle classe par rapport à telle autre en terme de résultats de coloriage de corpus.

On pourra également chercher à proposer à l'utilisateur de compléter les classes extraites et structurées à l'aide de ThemeEditor et Anadia avec des bases de données lexicales, comme par exemple le système développé par (Ploux et Victorri, 1998), pour y ajouter notamment des relations d'homonymie et de synonymie.

Références

Beust P. (1998). Contribution à un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines, Thèse de l'Université de Caen.

Beust P. (2000). Pour une modélisation de la référenciation dans le dialogue homme-machine, Journée d'étude de l'ATALA : Référence et T.A.L., 18/11/2000.

Daille B. (1995). ACABIT : une maquette d'aide à la construction automatique de banques terminologiques monolingues ou bilingues, IVe journée scientifiques de l'AUPELF-UREF : lexicomatique et Dictionnairiques, Lyon.

Enguehard C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle ILN'93*, pages 373-384, Nantes.

Faure D. (2000). Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM, Thèse de Doctorat Université de Paris Sud.

Giguet E. (1998). Méthode pour l'analyse automatique de structures formelles sur documents multilingues. Thèse de l'Université de Caen.

Greimas A. J. (1983). *Du sens II*, Essais sémiotiques, Paris, Editions du Seuil.

Nicolle A., Beust P. et Perlerin V. (2002). Un analogue de la mémoire pour un agent logiciel interactif, *In Cognito*, à paraître.

Perlerin V. (2000). Catégorisation lexicale et recherche documentaire, Rapport de DEA «intelligence artificielle et algorithmique », Université de Caen.

Perlerin V. (2001). La recherche documentaire : une activité langagière, *Récital, TALN2001*, pages 469-478.

Pichon R. et Sébillot P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience, *TALN1999*, pages 279-288.

Ploux S. et Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes, *TAL* vol.(39:1).

Rastier F. (1987). *Sémantique interprétative*, Paris, Presses Universitaires de France.

Rastier F. (1991). *Sémantique et recherches cognitives*, Paris, Presses Universitaires de France.

Rastier F., Cavazza M. et Abeillé A. (1994). *Sémantique pour l'analyse*, Paris, Masson.

Saussure F. de (1915). *Cours de linguistique générale*, Paris 1986, Ed. Mauro-Payot.

Tanguy L. (1997). Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative, Thèse de l'Université de Rennes I.

Thlivitis T. (1998). Sémantique interprétative Intertextuelle : assistance anthropocentrée à la compréhension des textes, Thèse d'informatique de l'Université de Rennes I.

Victorri B. et Fuchs C. (1996). *La polysémie*, Paris, Hermes.

Zipf G.K. (1949). *Human Behavior and the Principle of least effort : an introduction to Human Ecology*, Mass: Addison-Wesley, Reading.