# Analyse conjointe texte et image pour la caractérisation de films d'animation

Païs Grégory

THESE

présentée et soutenue publiquement le 06 avril 2010 à Annecy-le-Vieux

en vue du Doctorat de

l'UNIVERSITE DE SAVOIE

mention STIC Traitement de l'Information par

**M. Grégory PAÏS**

---

# Analyse conjointe texte et image pour la caractérisation de films d'animation.

---

Thèse préparée au LISTIC et encadrée par : Pr Patrick LAMBERT
Dr Daniel BEAUCHENE

---

COMPOSITION DU JURY

| | |
|---|---|
| Président : | Pr Matthieu CORD, UPMC-PARIS VI, Paris |
| Rapporteurs : | Pr Christophe DUCOTTET, Univ. Jean Monnet, Saint-Etienne |
| | Pr Sylvie DESPRÉS, Univ. Paris 13, Paris |
| Examinateurs : | Pr Matthieu CORD, UPMC-PARIS VI, Paris |
| | Dr Ladjel BELLATRECHE, ENSMA, Poitiers |
| | Dr Françoise DELOULE, Univ. Savoie, Chambéry |

**Résumé :** Le développement rapide des nouvelles technologies de l'information a provoqué ces dernières années une augmentation considérable de la masse de données à disposition de l'utilisateur. Afin d'exploiter de manière rationnelle et efficace l'ensemble de ces données la solution passe par l'indexation de ces documents multimédia. C'est dans ce contexte que ce situe cette thèse et plus spécifiquement dans celui de l'indexation d'une base numérique de films d'animation, telle que celle mise en place par la CITIA (Cité de l'image en mouvement).

L'objectif principal de cette thèse est de proposer une méthodologie permettant de prendre en compte des informations issues de l'analyse de l'image et celles issues des péri-textes (synopsis, critiques, analyses, etc.). Ces deux sources d'information sont de niveau sémantique très différent et leur utilisation conjointe permet une caractérisation riche et sémantique des séquences vidéo. L'extraction automatique de descripteurs images est abordée dans ces travaux à travers la caractérisation des couleurs et de l'activité du film. L'analyse automatique des synopsis permet quant à elle de caractériser la thématique du film et permet, grâce au scénario actanciel, la caractérisation de l'action de la séquence. Finalement ces informations sont utilisées conjointement pour retrouver et décrire localement les passages d'action et permettent d'obtenir l'atmosphère du film grâce à leur fusion floue.

**Abstract :** These last years, the fast development of new technologies allows digital media collections and the circulation of these data growing in size and number. However, the exploitation of these data remains a whole problem and creates a strong requirement for efficient tools to manipulate it. The current trend is in search of automatic indexing technique, based on semantic document contents. In the context of Annecy International Animation Film Festival the animated movie characterization presented in our works consists in information fusion between information contained in the animated movie images, textual information extracted from festival registration form and the expert knowledge. This information fusion uses color statistics and activity measure extracted from automatic image sequence analysis and uses textual description and emotion measure from automatic textual analysis. Two characterizations are provided from this information fusion. A first global characterization consists in a dramatic emotion classification and a second local characterization consists in the time-localized action description from the actant scenario.

**Mots clefs :** Caractérisation de films, cinéma d'animation, analyse d'images, analyse de textes, Extraction d'information, scénario actanciel, fusion d'information, système de fusion flou.

# Remerciements

Cette thèse a été menée au sein du Laboratoire d'Informatique, Systèmes et Traitement de l'Information et de la Connaissance (LISTIC) de l'Université de Savoie et a été co-financée par l'Assemblée des Pays de Savoie (APS), le LISTIC et Polytech'Savoie. Mes premiers remerciements vont donc naturellement à ces institutions.

Je remercie madame Sylvie Després et monsieur Christophe Ducottet d'avoir bien voulu rapporter mes travaux ainsi que les membres du jury monsieur Matthieu Cord et monsieur Ladjel Bellatreche pour leurs remarques et suggestions concernant ce manuscrit et plus généralement sur mes travaux de thèse.

La soutenance d'une thèse est un événement unique qui marque la fin d'une expérience qui fut très enrichissante. Ainsi je tiens à remercier chaleureusement madame Françoise Delloule, messieurs Daniel Beauchêne et Patrick Lambert qui m'ont accompagné tout au long de ces années. J'ai une sincère gratitude pour ces trois chercheurs de communautés scientifiques différentes auprès desquels j'ai beaucoup appris.

Je souhaite remercier Philippe Bolon et plus généralement l'ensemble du laboratoire de m'avoir accueilli au sein du LISTIC. Un grand merci en particulier à Joëlle et Samia notre secrétariat de choc pour leur disponibilité, leur écoute et leur gentillesse.

Un merci tout particulier à mes collègues de bureau Amory Bisserier, Olivier Passalacqua, Florent Martin et Nabile Fakhfakh pour leur soutien, leur sympathie, et les grands moments musicaux où les extravagances vocales étaient au rendez-vous. Un grand merci également aux doctorants et personnel technique Azadeh, Yajing, Andreea, Renaud, Sylvain, Alain, Abdellah, Fabien et Sébastien.

Merci mille fois à mes amis et tout ceux qui m'ont soutenu par votre présence ou vos messages d'encouragement le jour de la soutenance Céline, Amandine, Baba, Émilie, Pat, Juju, Pim's, Soso, Christelle, Anne-So, Patoune, Guillaine, Max, Alex, Delf, Flo, Yo et j'en oublie... Merci également à Not'in Game Gospel mon échappatoire musicale pour l'émotion et la richesse humaine que vous m'apportez. Mes derniers remerciements vont tout naturellement à ceux qui partage ma vie. Je pense bien sûr à ma famille qui m'a soutenu et qui m'a permis d'aller jusqu'au bout de ce projet de thèse.

***Il faut avoir une musique en soi pour faire danser le monde.***
*Nietzsche Friedrich Wilhelm*

# Table des matières

# Liste des acronymes

**CITIA** Cité de l'Image en Mouvement

**FIFA** Festival International du Film d'Animation d'Annecy

**MIFA** Marché International du Film d'Animation

**CICA** Centre International du Cinéma d'Animation

**LISTIC** Laboratoire d'Informatique Systèmes, Traitement de l'Information et de la Connaissance

**SCC** Short Color Change

**NBS** National Bureau of Standards

**ISCC** Inter-Society Color Cuncil

**CMC** Color Measurement Committee

**AaA** Algorithme à Accumulation

**CAH** Classification Ascendante Hiérarchique

**EI** Extraction d'Information

**NE** Named Entity

**CO** Coreference resolution

**TE** Template Element

**TR** Template Relation

**ST** Scenario Template

**MUC** Message Understanding Conference

**LG** Link Grammar

**TAL** Traitement Automatique de la Langue

**AFC** Analyse Factorielle des Correspondances

**OCR** Optical Character Recognition

**ASR** Automatic Speech Recognition

**SVM** Support Vector Machine

**FFT** Fast Fourier Transform

**STFT** Short-Time Fourier Transform

**ADT** Analyse de Données Textuelles

**SMO** Sequential Minimal Optimization

**MLP** Multi-Layer Perceptron

# Première partie

# Introduction

Chapitre 1

# Le contexte général : les systèmes d'indexation

**Résumé :** Dans ce chapitre nous abordons les problématiques d'indexation de documents multimédias qui passent par les problématiques de caractérisation de ces documents. Cet état de l'art du domaine présente les différentes solutions habituellement mises en œuvre dans les problèmes de caractérisation et d'indexation des documents contenant aussi bien des images, du son, des textes que de vidéos.

## 1.1 Présentation du contexte général

La masse de données multimédia personnelles ou collaboratives est en très forte augmentation depuis quelques années. Le stockage et la circulation de ces données, informations ou connaissances, sont facilités par le développement rapide des nouvelles technologies de l'information de ces dernières années. Cependant, l'exploitation rationnelle et efficace de ces grandes masses de données reste un problème entier. Ainsi l'identification des informations pertinentes d'un document passe par une opération d'indexation. Cette opération consiste à analyser le contenu de ce document et à le transcrire dans un langage documentaire. Cette normalisation et codification du contenu des documents reposent sur des index qui permettent de classer un document parmi un ensemble de documents d'une collection donnée et facilitent in fine la recherche de ce document pour l'utilisateur. Traditionnellement l'indexation par mots-clés, qui s'appuie sur une information externe (de type liste de mots), est lourde à mettre en œuvre et manque parfois d'efficacité pour des documents vidéo (par exemple, retrouver des passages spécifiques dans une vidéo indexée globalement). Cette information externe est le plus souvent issue d'un opérateur humain qui analyse et catégorise le document suivant son contenu et l'interprétation qu'il en fait. Cette approche est quasiment infaisable lorsque l'on doit indexer de grandes quantités d'informations (pages web ou images de vidéos par exemple). En effet, la lecture et/ou l'interprétation des documents est un processus cognitif complexe qui prend un certain temps. En revanche, ces index ont l'avantage d'être d'un haut niveau sémantique puisque directement issus et formalisés par l'homme, mais souffrent quelquefois d'une certaine hétérogénéité, car ils proviennent d'une interprétation humaine. Depuis les années 90 la tendance est à la recherche de techniques d'indexation automatiques ou semi-automatiques, basées sur le contenu et la sémantique. Ce problème est un véritable

défi et se décline suivant plusieurs aspects. Les informations extraites à partir des documents doivent permettre une indexation fiable et pertinente. La modélisation, la représentation et l'organisation de cette information doivent être souples et efficaces. De plus, l'interaction homme-machine nécessaire à l'utilisation et la consultation de ces bases de données multimédias (dont la nature est variable : textuelle ou/et vidéo ou/et image) est une problématique importante qui n'a été prise en compte que récemment. Cette thématique de recherche émergente est désormais considérée comme une thématique prioritaire et des efforts importants sont déployés à travers le monde pour apporter des solutions fonctionnelles dont les retombées techniques et économiques sont considérables (chiffre d'affaire de *Google* en 2008 $\simeq$ 6 milliards de \$). Trois grands champs se dégagent dans la tâche de recherche de documents multimédia :

- **L'extraction de descripteurs et la caractérisation des documents**. Ce champ consiste en l'extraction de descripteurs de haut niveau, basés sur le document et sur des connaissances universelles ou spécifiques au document ou au domaine. Ces descriptions sont issues directement de l'information contenue sur le support documentaire et transformées en caractéristiques de haut niveau sémantique.

- **L'organisation et la gestion des documents**. Ce champ consiste à organiser et modéliser les descriptions/index des documents pour permettre la recherche et la visualisation de ces documents. Les techniques utilisées sont basées sur les modèles de connaissances de la gestion documentaire et du web sémantique. Afin d'exploiter efficacement ou d'aider à la génération de ces index, annotations ou méta-données, ces connaissances sont généralement issues des modèles du domaine, du savoir-faire des auteurs et/ou des pratiques et besoins des utilisateurs.

- **La recherche et visualisation des documents**. Ce champ consiste à retrouver à partir d'une requête formulée par l'utilisateur, les documents ou fragments documentaires préalablement indexés et permettre une visualisation et une navigation dans le corpus documentaire. Dans le cas de vidéos par exemple, la visualisation des documents est souvent aidée par des outils de synthèse permettant de créer des résumés des documents originaux afin d'accélérer la navigation dans la base. Les techniques utilisées sont basées sur des modèles de connaissances (modèles de domaine, pratiques des utilisateurs, etc.) et sur les contraintes liées aux infrastructures et aux supports de visualisation, etc.

Dans cette thèse nous nous intéressons principalement au premier champ énoncé ci-dessus, c'est-à-dire à l'**obtention** de descripteurs issus du/des signal(aux) et à leurs **transformations** pour obtenir des descripteurs de plus haut niveau sémantique.

## 1.2 Les systèmes d'indexation

Nous venons de voir que pour rechercher et manipuler les documents multimédia, les index décrivant leur contenu doivent être riches et aussi complets que possible. Pour cela, la caractérisation puis l'indexation des données peuvent prendre deux formes principales : *l'annotation manuelle* ou *l'annotation automatique*. Qu'elle soit manuelle ou automatique, le coût (temps d'analyse, complexité algorithmique, nombre et pertinence des index, souplesse, etc.) de l'annotation de contenu est directement lié au niveau de détail désiré, ce qui

définit la granularité de l'indexation [Faudemay *et al.*, 1998, Ramesh *et al.*, 2002]. En effet, pour un niveau de détail élevé il faut une analyse plus importante du contenu des données. L'annotation humaine a l'avantage d'offrir des index de haut niveau sémantique mais elle est lourde et demande beaucoup de ressources et de temps humain. Les méthodes automatiques assistées par ordinateur sont beaucoup plus rapides car elles ne demandent pas ou peu d'intervention humaine. Cependant elles ne sont pas capables à l'heure actuelle de fournir les mêmes informations que les méthodes manuelles. Ce constat, établi depuis longtemps pour les documents textuels [Salton, 1968, Anderson et Pérez-Carballo, 2001], est d'autant plus vrai pour les documents multimédia (image, musique, vidéo) qui ne bénéficient pas d'une tradition aussi longue et dont l'indexation nécessite une interprétation du contenu qui passe par l'utilisation de techniques d'analyse avancées basées sur l'intelligence artificielle. Cette interprétation qui reste un processus cognitif complexe souffre en l'état des techniques de ne pas apporter la richesse informationnelle (niveau sémantique) attendue par l'homme. Cet écart entre l'information extraite automatiquement par des algorithmes et l'information issue de l'interprétation humaine est le véritable verrou technologique dans la tâche d'indexation multimédia et porte le nom de **fossé sémantique** ou "semantic gap". Ce fossé est particulièrement important lorsque les documents à indexer sont des images. Le moteur de recherche *Google* a beau être leader dans son domaine, il n'en reste pas moins qu'il reste confronté à cet épineux problème qu'est l'indexation des images sur Internet. Dans ce moteur de recherche, les images sont indexées à partir du texte entourant l'image à l'intérieur de la page web. Ainsi le système ne tient pas compte du contenu sémantique propre à l'image. Par exemple, on voit sur la figure 1.1 que dans les image 11,12,15 et 17 (en partant d'en haut à gauche) montrent une voiture et une machine à bois : ces images ont été abusivement indexées par le mot clef "WAGON" car la référence textuelle (le nom du fichier) associée, contient le mot "wagon", le contenu de l'image ayant été complètement occulté.

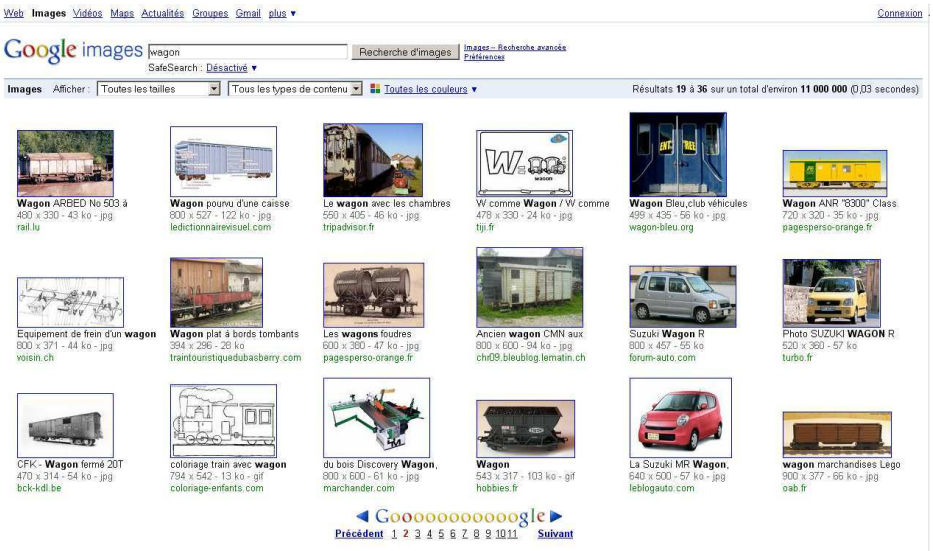



FIGURE 1.1 – Résultat proposé par Google-Image pour la requête "wagon"

Ce moteur de recherche permet cependant un filtrage basé sur des attributs bas niveau issus directement de l'image : un filtrage suivant la couleur dominante peut être réalisé ou bien encore une reconnaissance du type d'image (dessin, photo, clip-art) ou une détection des

visages (options accessibles à partir du bandeau supérieur sur la figure 1.1). Nous voyons sur cet exemple de moteur de recherche que les informations issues du document image restent pauvres en sémantique et demeurent bien loin des informations de contenu attendues par l'utilisateur. Le travail d'indexation des images est complexe et de longue haleine. Une nouvelle piste, mais qui utilise encore l'intervention humaine, est explorée par *Google* qui fait appel à ses internautes pour l'aider dans l'indexation et le référencement des images sur Internet en leur demandant d'attribuer eux-même des mots clefs par l'intermédiaire d'un jeu interactif [1].

A travers cet exemple on voit toute la difficulté du travail d'indexation qui reste une thématique de recherche essentielle. Ainsi pour [Snoek et Worring, 2005] trois questions se posent. La première est liée à la granularité des indexations et s'exprime ainsi : **"Quoi indexer ?"** : par exemple, le document en entier, les grandes parties ou des parties plus fines. La deuxième est liée à la nature des index et s'exprime ainsi : **"Quels index ?"** : par exemple, le nom des joueurs dans un match de football, leurs positions et déplacements dans le temps, etc. La troisième est liée aux sources ou modalités et à leur analyse et s'exprime ainsi : **"Comment indexer ?"** : par exemple, utiliser un classifieur statistique appliqué au contenu auditif seulement ou utiliser un algorithme de reconnaissance de visages. Pour cette dernière question, de nombreuses solutions ont été développées depuis une vingtaine d'années, basées souvent sur une approche uni-modale, c'est-à-dire sur l'utilisation d'une seule source d'information pour caractériser le document. Si nous reprenons l'exemple de la figure 1.1, la modalité exploitée est le texte. Cette uni-modalité se retrouve dans les premiers systèmes d'indexation de contenu qui sont basés sur la similarité de descripteurs bas niveau ("features") extraits d'une des modalités du média. Ces descripteurs sont généralement extraits de la modalité la plus caractéristique du document traité (descripteurs image dans le cas de photo, descripteurs audio dans le cas de musique, etc.).

### 1.2.1 Les systèmes d'indexation image

Dans les systèmes d'indexation de bases d'images statiques ou CBIR ("Content-Based Image Retrieval"), l'analyse du contenu des images se décline suivant trois axes principaux : *la couleur*, *la forme* et *la texture* [Smeulders *et al.*, 2000, Liu *et al.*, 2007].

**L'analyse des couleurs** est une caractéristique fondamentale dans le système visuel humain. De ce fait c'est une des directions les plus utilisées dans les algorithmes d'analyse d'images. Les couleurs sont analysées en utilisant différents espaces, en commençant par le classique espace RVB et en passant à des espaces plus complexes comme par exemple les espaces perceptuels (HSV, Lab, etc.).

**L'analyse des formes** utilise les propriétés géométriques des objets contenus dans l'image pour caractériser la scène. Ceci demande en général la détection préalable des objets, le plus souvent par des techniques de segmentation par approche "contours" ou "régions". Ces caractéristiques ne doivent pas dépendre du point de vue sous lequel ces objets sont observés. Différents descripteurs de formes, invariants aux transformations géométriques de l'image [Rivlin et Weiss, 1995], sont proposés et permettent également de solutionner le problème d'occlusion entre différents objets lié à la projection de l'espace réel 3D dans l'espace 2D de l'image [Schmid et Mohr, 1997].

1. http://images.google.com/imagelabeler

**L'analyse des textures** est également très utilisée car ces informations permettent de caractériser les propriétés des matériaux présents dans l'image. La classification de texture trouve par exemple des applications dans la recherche d'images [Gimel'Farb et Jain, 1996] ou l'analyse d'images médicales. Les démarches existantes se divisent en trois approches fondamentales [Liu *et al.*, 2009] : les approches statistiques, les approches structurelles et les approches spectrales. Parmi ces trois approches de nombreux algorithmes ont été proposés comme l'utilisation des moments statistiques dans [Avilés-Cruz *et al.*, 2005] ou l'utilisation de modèle markovien dans [Choi et Baraniuk, 2001] ou encore l'utilisation d'ondelettes dans [Pothos *et al.*, 2007].

Pour [Liu *et al.*, 2007], la *position spatiale* est aussi un descripteur bas niveau souvent utilisé dans les systèmes de caractérisation d'images. Par exemple, le ciel et la mer peuvent avoir des descripteurs couleur et de texture assez similaires alors que leurs dispositions spatiales sont différentes (partie supérieure de l'image pour le ciel et partie inférieure pour la mer).

Les systèmes d'indexation s'appuient souvent sur une étape de catégorisation. Les méthodes de catégorisation peuvent être classées en deux catégories [Pujol, 2009] : une caractérisation globale qui consiste à caractériser l'image dans sa totalité (paysage naturel, couchers/levers de soleil, etc ) et une caractérisation locale qui consiste à caractériser les éléments composant l'image (arbre, mer, soleil, etc). Ces approches tentent de répondre au problème du passage à la sémantique par l'utilisation de classifieurs (et de leurs combinaisons) afin d'associer un concept connu à un ensemble de valeurs des descripteurs images. Ces méthodes utilisent le plus souvent des algorithmes d'apprentissage supervisé comme les Support Vector Machine (SVM), les réseaux de neurones ou des approches probabilistes basées sur des modèles comme les "Mixtures de Gaussiennes" (GMM) ou les chaines de Markov cachées (HMM). Ces dernières années sont apparues des méthodes performantes, dites par *sacs de mots* ("bag of features"), utilisées pour catégoriser des images. Ces méthodes sont inspirées de la linguistique où les documents textuels sont caractérisés par un ensemble de mots ("sac de mots"), issus d'un dictionnaire, où l'ordre des mots dans le regroupement ("sac") est sans importance. En vision, cette méthode consiste à modéliser une image par un ensemble de "mots" qui sont en réalité une simple distribution de caractéristiques locales (texture, couleur, etc) extraites de régions d'intérêts. Un des premiers travaux sur ce principe a permis dans [Ullman *et al.*, 2001] de retrouver des voitures et des visages à partir de fragments d'image. De nombreux travaux [Lazebnik *et al.*, 2006, Larlus, 2008] se sont également inspirés de cette approche permettant dans [Lazebnik *et al.*, 2006] de retrouver des scènes plus ou moins complexes (campagne, ville, forêt, chambre, etc) ou dans [Van de Sande *et al.*, 2008] de retrouver de nombreux concepts issus du *"PASCAL Visual Object Challenge"* et du *"Mediamill Challenge"*. Pour un état de l'art en recherche d'image voir [Datta *et al.*, 2008].

### 1.2.2 Les systèmes d'indexation audio

On retrouve également cette uni-modalité dans les systèmes d'indexation audio. Dans de tels systèmes d'indexation, l'analyse du contenu des documents sonores est développée selon deux axes principaux : *la représentation temps-amplitude du signal*, ainsi que *la représentation spectrale du signal* [Lu, 2001]. Parmi les descripteurs temporels les plus couramment

utilisés on peut citer :

**L'énergie moyenne** : elle permet de caractériser le volume du signal sonore. Son calcul peut se faire de plusieurs manières et permet de discriminer la parole de la musique car la parole présente généralement plus de variations que la musique.

**Le ZCR** : c'est le taux de passages à zéro de la forme d'onde temporelle ("Zero Crossing Rate" ou ZCR). Il caractérise la fréquence de changements de signe du signal. Cela permet de détecter les signaux de parole par les brusques variations de son profil temporel.

D'autres descripteurs sont calculés dans le domaine fréquentiel :

**Le spectre fréquentiel** permet de caractériser le signal par la distribution des fréquences. Son calcul se fait classiquement à partir du signal temporel en utilisant la transformée de Fourier et de ses nombreuses variantes (Fast Fourier Transform (FFT), Short-Time Fourier Transform (STFT), Ondelettes, etc). L'analyse du spectrogramme (diagramme associant à chaque instant t d'un signal son spectre fréquentiel) permet d'identifier des sons, comme les timbres des instruments musicaux ou la parole. On peut citer un certain nombre de mesures caractérisant le spectre fréquentiel comme la mesure d'asymétrie ("Skewness"), le calcul du coefficient d'aplatissement ("Kurtosis") ou encore le calcul des MFCCs ("Mel-Frequency Cepstral Coefficients").

**L'harmonicité** accompagne souvent la mesure de la fréquence fondamentale. Le degré d'harmonicité du signal permet de mesurer la richesse d'un son en harmonique (multiples de la fréquence fondamentale). Un son parfaitement harmonique est un son dont les raies spectrales sont situées à des fréquences multiples entières de la fréquence fondamentale. En outre, ce descripteur est un bon indicateur dans la classification des timbres des instruments de musique.

**Le vecteur de chroma** ou *"Chroma features"* est un puissant descripteur pour représenter les signaux musicaux. Le spectre du signal audio est projeté sur l'échelle chromatique (composée de 12 demi-tons dans une octave) créant ainsi un vecteur de Chroma. Ce vecteur intègre l'énergie dans toutes les bandes correspondant à chacun des douze degrés de l'échelle chromatique de la gamme musicale. De plus, les notes séparées d'exactement une octave étant perçues comme semblables, connaître la distribution du chroma même sans fréquence absolue (c'est-à-dire l'octave originale) peut fournir des informations musicales utiles sur le morceau et permet de mesurer la similitude musicale perçue.

De nombreuses solutions ont été développées à partir de ces indicateurs et les mélomanes peuvent par exemple retrouver des morceaux de musique par similarité rythmique [Foote, 1999] ou par similarité des séquences d'accords [Hanna *et al.*, 2009]. Dans [Essid, 2005], ces descripteurs permettent de retrouver les instruments de musique utilisés dans les morceaux musicaux. Ils permettent aussi de détecter des pleurs dans [Petridis et Pantic, 2008], de retrouver des événements dans les vidéo sportives (applaudissements, frappe dans la balle, sifflement, etc.), ou de reconnaître les émotions transmises dans la musique [Trohidis *et al.*, 2008] et la parole [Xiao *et al.*, 2009]. Pour plus de détails voir les états de l'art de [Scaringella *et al.*, 2006] et de [Orio, 2006].

### 1.2.3 Les systèmes d'indexation de séquences d'images

L'arrivée et la diffusion de la vidéo numérique ont orienté les systèmes d'indexation vers *les séquences d'images*. Les systèmes d'indexation des séquences d'images ou CBISR ("Content-Based Image Sequence Retrieval") sont à la base l'extension temporelle des systèmes CBIR. Dans ce cas, le traitement n'est pas effectué sur des images statiques indépendantes les unes des autres, mais sur des séquences qui sont des suites temporelles d'images ou des images en mouvement. Le premier problème qui se pose est *le volume des données*. A une cadence de 25 images par seconde, une séquence d'images de 10 minutes contient *15000 images*. Un film à lui seul est ainsi équivalent, du point de vue de la taille, à une base contenant plusieurs dizaines de milliers d'images, avec bien sûr une forte redondance de l'information entre les images. D'autre part, à l'information spatiale fournie par l'image s'ajoute une nouvelle information à traiter : *l'information temporelle*. Si dans un système CBIR deux images qui contiennent les mêmes objets sont considérées comme similaires du point de vue de leur contenu, dans un système CBISR deux séquences d'images contenant les mêmes objets peuvent avoir des contenus très différents si l'on prend en compte l'aspect temporel. Ainsi, le comportement des objets et l'évolution temporelle de la scène sont des informations essentielles pour la compréhension du contenu des séquences et donc pour la tâche d'indexation.

De nombreux travaux portant sur l'analyse du mouvement dans les séquences d'images ont été entrepris durant cette décennie. Les caractéristiques mesurées par l'analyse de mouvement ("motion analysis") découlent généralement d'approches **spatio-temporelles**. Cela permet par exemple, d'extraire des caractéristiques sur la trajectoire des objets, ou sur les mouvements de caméra (zoom, travelling, rotation, ...). Deux grandes directions d'analyse sont définies dans [Jeannin et Divakaran, 2001], d'une part l'analyse du *mouvement global* de la caméra et d'autre part l'analyse locale du *mouvement des objets*

**Le mouvement global**. L'analyse du *mouvement global* est effectuée au niveau des plans vidéo ou de groupes d'images. Une première information extraite est le *mouvement de la caméra*. Les déplacements particuliers de la caméra sont déterminés parmi tout un ensemble de mouvements possibles. Typiquement les informations retenues pour détecter un mouvement spécifique sont la direction et l'amplitude du mouvement, sa position dans la séquence et sa durée. De nombreuses approches ont été explorées [Bouthemy *et al.*, 1999, Duan *et al.*, 2004] comme l'analyse des vecteurs de mouvement ("motion vector")[2] dont l'analyse des directions permet de déterminer le mouvement d'une caméra [Zhang *et al.*, 1995]. Cette information de mouvement permet de localiser les passages importants de la séquence comme par exemple le fait de focaliser l'attention des spectateurs (arrêt sur une scène précise, puis zoom sur le visage d'un personnage). Une seconde information souvent exploitée est *l'intensité du mouvement*. Cette grandeur qui mesure globalement l'amplitude du mouvement dans les images est un bon indicateur de l'activité de la scène.

**Le mouvement local**. La deuxième direction d'analyse est la caractérisation du *mouvement local ou mouvement des objets* qui n'affecte que certaines régions de l'image. Ces mesures de déplacement sont généralement effectuées au niveau du pixel et utilisent com-

2. vecteur 2D dans l'espace image représentant le déplacement d'un même bloc de pixels entre deux images. Il permet de passer des coordonnées du bloc dans l'image de référence à l'instant t, aux coordonnées du même bloc dans l'image à un temps t+1.

munément une modélisation du mouvement permettant de retrouver dans la séquence des déplacements similaires. Ces analyses permettent de retrouver la *trajectoire* des objets, définie par l'évolution temporelle de certains points d'intérêt de l'objet, comme par exemple le centre de gravité ou certains points de contour [Panagiotakis *et al.*, 2006]. Parmi les méthodes souvent rencontrées dans la littérature, on peut citer le flot optique ("optical flow") dont le calcul consiste à extraire un champ de vitesses dense à partir d'une séquence d'images en faisant l'hypothèse que l'intensité (ou la couleur) est conservée au cours du déplacement [Quénot, 1996]. On peut noter que le flot optique, bien que ce soit une mesure locale, peut-être utilisé pour obtenir une mesure globale du mouvement dans la séquence. D'autres méthodes de détection consistent à détecter dans le domaine spatial et temporel les points d'intérêts ("interest points") dont les valeurs image ont une variation locale significative à la fois dans l'espace mais aussi dans le temps [Laptev, 2005]. Pour un état de l'art sur les techniques d'analyse du mouvement des objets on pourra se rapporter à [Koprinska et Carrato, 2001] [Smith *et al.*, 2004] ou [Trucco et Plakas, 2006].

Les séquences d'images sont très souvent associées à une bande son (et plus rarement à du texte) pour former des documents vidéo. Bien entendu, ces informations (image, audio, texte) ne sont pas indépendantes les uns des autres. Il y a synchronisation de celles-ci et un certain nombre de liens sémantiques existent entre elles. Par exemple lorsqu'un joueur de football est dans une action et qu'il marque un but, l'enthousiasme du public et des commentateurs sportifs est mesurable sur la bande son [Leonardi *et al.*, 2003]. Cependant ces liens ne sont pas toujours triviaux et peuvent devenir subjectifs dans le cas de films artistiques, films auxquels nous allons nous intéresser plus spécifiquement dans la suite.

## 1.3 Les systèmes d'indexation de films

Dans ce paragraphe, nous nous limitons à l'indexation des vidéos particulières que sont les films. Un film est une œuvre produite par un auteur, et se distingue donc des autres vidéos, les vidéos de surveillance par exemple. En effet, un film est conçu dans un environnement de production et est le résultat d'un projet artistique voulu par son auteur. Pour exprimer cette idée, l'auteur utilise généralement plusieurs modalités :

- **La modalité visuelle :** c'est la mise en scène, c'est-à-dire tout ce qui est naturellement ou artificiellement créé et que le spectateur peut voir.

- **La modalité audio :** c'est tout ce que le spectateur peut entendre, c'est-à-dire la parole, la musique, les sons ambiants.

- **La modalité textuelle :** c'est tout ce que le spectateur peut lire. C'est par exemple le texte se superposant aux images, mais cela peut également être des *péri-textes*, c'est-à-dire des textes qui parlent du film (résumé, script, sous titrage ...).

L'exploitation conjointe de ces modalités à des fins de caractérisation du document vidéo semble de ce fait naturelle. Comme nous l'avons déjà évoqué, la caractérisation dépend du niveau de détails désiré. Pour les documents vidéo cette granularité peut s'exprimer suivant cinq niveaux (voir figure 1.2). Chaque niveau représente une (des) unité(s) multimodale(s) dont le contenu sémantique est homogène [Davenport *et al.*, 1991].

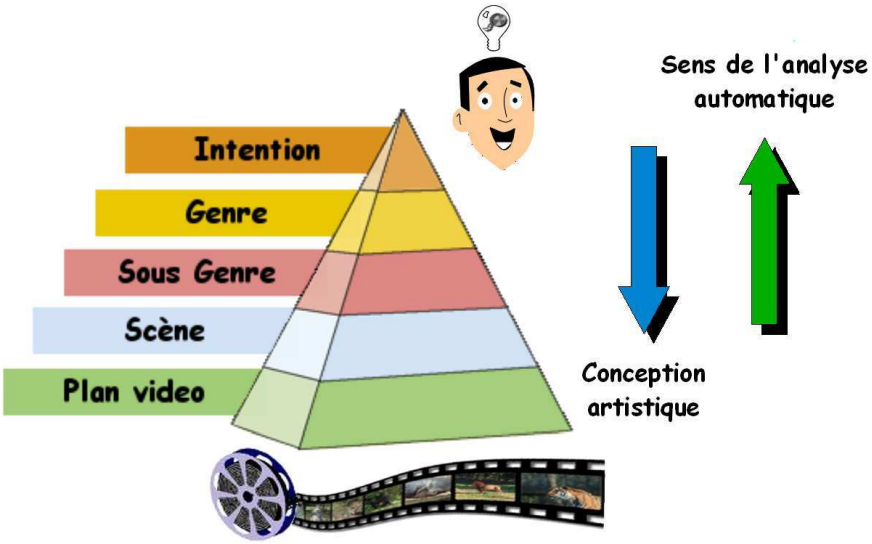



FIGURE 1.2 – Hiérarchie des différents niveaux sémantiques composant le document vidéo

- **L'intention** c'est-à-dire pourquoi le document existe (divertissement, information, communication, etc.).

- **Le genre** est en fait le style du document (long-métrage, documentaire, film publicitaire, etc.).

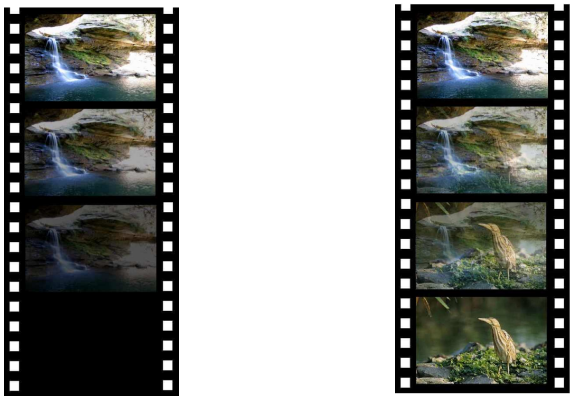
- **Le sous genre** est un sous-ensemble du genre dans lequel les documents vidéo partagent le même contenu sémantique (films d'horreur, d'aventure, policier, etc.).

  Les niveaux suivants découpent le document vidéo en segments basés sur l'homogénéité de la modalité ou du contenu, et correspondent donc à une décomposition du support.

- **L'unité logique ou scène** comprend un ensemble de plans qui sont liés du point de vue sémantique. Le contenu d'une scène doit respecter la règle des trois unités, comme dans le théâtre classique du XVII$^{e}$ siècle : unité de lieu, unité de temps et unité d'action [Corridoni et Del Bimbo, 1995]. C'est par exemple une course poursuite avec les forces de police dans un film d'action.

- **L'événement ou plan vidéo** est un court segment vidéo (événement nommé) dont le contenu ne change pas dans le temps. La séquence d'images ainsi obtenue présente une continuité visuelle. Cela peut être le moment où une voiture poursuivante fait un tonneau ou les images montrant une explosion dans une scène de course poursuite d'un film d'action. L'auteur utilise trois éléments pour construire un plan vidéo : le cadre qui est le lieu et le temps de l'événement, les objets statiques ou dynamiques qui sont des entités inanimées (dans le sens de "non vivant") et enfin les personnages qui sont des personnes ou des animaux.

Dans le cas d'un film, l'auteur, lors de la création de son document visuel, part d'une intention artistique et utilise une ou plusieurs modalités pour construire son film. Chacune de ces modalités est une séquence temporelle constituée d'éléments fondamentaux dont la nature est intrinsèquement liée au support de l'information. La modalité visuelle par exemple, est une succession temporelle d'images dont l'élément fondamental est l'image. De même, la modalité audio est constituée d'éléments fondamentaux qui sont des échantillons audio d'une courte durée (notes dans le cas de la musique). Les différents caractères forment les unités fondamentales pour la modalité textuelle. Ainsi, l'agrégation de ces éléments fondamentaux forme le plan vidéo. La création prend alors tout son sens lorsque l'auteur concatène ces différents plans à partir de chacune des modalités, pour former un ensemble cohérent (une scène) qui constituera in fine le document tout entier. Cette juxtaposition se fait grâce à l'utilisation de transitions dont l'utilisation est importante pour obtenir une continuité ou une discontinuité, visuelle ou sonore [Lienhart, 2001]. Dans les films, on retrouve principalement trois types de transition :



(a) fondu au noir "fade out" (b) fondu enchaîné "dissolve"

FIGURE 1.3 – Transitions graduelles pour la modalité image

- Abrupte de type **"cut"** : c'est le passage d'un plan à un autre par simple juxtaposition de ceux-ci (transition abrupte entre deux images pour la modalité image ou utilisation d'un silence pour obtenir une transition franche pour la modalité audio [Boggs, 1996]).

- Graduelle de type **fondu** ou **"fade"** : les images s'assombrissent progressivement jusqu'à ce qu'elles deviennent entièrement noires (fondu à la fermeture "fade out" figure 1.3.a). Le plan suivant peut alors commencer par une image noire et s'éclaircir jusqu'à être normalement visible (fondu à l'ouverture "fade in"). Ou alors c'est la baisse du niveau sonore jusqu'au silence pour la modalité audio.

- Graduelle de type **fondu enchaîné** ou **"dissolve"** : cela consiste à superposer deux plans durant un court laps de temps, en diminuant la luminosité du premier tout en augmentant celle du second (dans le cas de la modalité image voir figure 1.3.b). Cela consiste aussi à baisser graduellement le volume d'un signal sonore tout en augmentant le volume d'un second.

Dans le texte, les transitions sont portées par les signes de ponctuation (espaces, points, guillemets, etc.) ou des effets graphiques (mise en couleur, en gras, etc.) [Perrot, 1980].

Lorsque l'on cherche à caractériser un document vidéo et plus particulièrement un film, une part importante de la caractérisation consiste à retrouver les différents niveaux sémantiques qui le composent (voir figure 1.2). Lorsque l'on ne dispose que du film, cette analyse automatique ne peut se faire que du bas vers le haut de la pyramide (l'intention de l'auteur n'est pas toujours une information disponible). La caractérisation du document doit donc commencer par la recherche de la structuration, sur les différentes modalités, de l'ensemble des informations attachées au document, et en particulier par la détection des plans vidéo. La connaissance de cette structuration (ou segmentation) permet ensuite de remonter aux scènes puis aux sous genres et genres pour arriver enfin à l'intention de l'auteur. Ainsi, un système d'analyse automatique doit, dans un premier temps, être capable de segmenter le document.

### 1.3.1 La segmentation des documents vidéo

La segmentation en plans est généralement obtenue à partir de l'analyse des modalités *image* et/ou *audio*. Habituellement, une seule de ces modalités (le plus souvent l'image) sert pour l'analyse automatique de la segmentation car, dans le cas des films, il y a habituellement synchronisation entre ces deux modalités. Notons que la segmentation des textes, lorsqu'ils existent, est généralement faite indépendamment des informations image et audio.

**Segmentation de la modalité image :** Plusieurs techniques permettent de retrouver les plans vidéo à partir des séquences d'images. Elles utilisent la détection des transitions vidéo et portent le nom de "shot boundary detection". La littérature abonde d'algorithmes de détection de transitions abruptes ("cut") fondés sur la comparaison des images successives de la séquence vidéo. Ils se basent sur la comparaison de pixels, de contours, de textures, de points d'intérêts, de blocs ou de vecteurs de mouvement. Les seuils nécessaires à ces comparaisons sont fixés manuellement ou dynamiquement Ces techniques peuvent être calculées directement depuis les images ou alors depuis le flux de données dans le cas de vidéo compressées (MPEG). On pourra se reporter aux états de l'art de [Aigrain *et al.*, 1996] et [Brunelli *et al.*, 1999] pour plus de détails. Cependant la comparaison image à image pour la détection de transition progressive est souvent insuffisante car les changements au sein des images successives ne sont pas toujours significatifs. Différentes techniques ont été mises au point pour palier ce problème dans [Zhang *et al.*, 1993, Arman *et al.*, 1993, Corridoni et Del Bimbo, 1995, Lu et Suganthan, 2004]. Pour plus de détails, voir un état de l'art dans [Cotsaces *et al.*, 2006].

**Segmentation de la modalité audio :** De nombreux travaux existent en ce qui concerne la segmentation d'un document à partir de la bande son [Carré et Philippe, 2000]. La détection de transitions brutales s'apparente à la détection des silences qui, dans [Patel et Sethi, 1996], est réalisée par l'analyse de l'énergie moyenne. Si la moyenne de l'énergie du signal, pour une fenêtre temporelle donnée, est inférieure à un seuil alors un silence de même longueur que la fenêtre est détecté. A partir du calcul des spectres et de l'extraction de certaines propriétés, Essid dans [Essid, 2005] segmente les passages musicaux. Pour de plus amples informations on pourra se reporter à l'état de l'art de [Carré et Philippe, 2000]).

**Segmentation de la modalité texte :** La segmentation d'un texte (chapitre, section, paragraphe, ...) passe par la détection des signes et marqueurs typographiques tels que :

les signes de ponctuation [Mourad, 1999], la cohésion lexicale [Choi et Baraniuk, 2001] ou la détection de la rupture de thème [Sitbon et Bellot, 2005]. Cependant ces techniques ne s'appliquent que sur des textes longs (plusieurs milliers de mots) et de tels textes ne sont généralement pas disponibles avec le document vidéo (sauf dans le cas de document sous titré où le fichier texte de sous titrage est disponible sur le support DVD). Cependant, même si à l'heure actuelle les films regroupent principalement les modalités image et audio, il n'est pas exclu d'intégrer cette modalité textuelle à l'avenir (avec le format MPEG7 par exemple). Ainsi, l'exploitation de la description textuelle (des images et de l'action du film par exemple) en synchronisation avec les autres modalités (comme dans un story board) pourrait être envisagée pour segmenter le film. Toutefois de tels travaux, exploitant la modalité textuelle pour segmenter le document vidéo, sont très peu fréquents (non trouvés dans la littérature consultée).

Une fois la structure du document obtenue, l'étape suivante consiste à analyser et caractériser les différents plans afin d'extraire une description, si possible sémantique, de leur contenu. La caractérisation de l'information contenue dans ces plans vidéo permet par la suite de retrouver les scènes et de remonter la pyramide. Après sa capacité de segmentation, un système d'analyse automatique doit donc être en mesure d'analyser et de décrire l'information contenue dans les plans vidéo.

### 1.3.2 La description du contenu

Afin d'accéder au contenu des plans vidéo, les techniques automatiques doivent être capables de retrouver et caractériser les objets et personnages constituant la scène en tirant profit des différentes sources d'information. Les approches d'extraction de contenu peuvent être regroupées en trois groupes : la détection de *scène* et de *concept*, la détection d'*objets* et la détection de *personnes*. Typiquement, les objets et personnes sont les éléments principaux que l'on retrouve dans les plans vidéo. Leur apparence est voulue par l'auteur en utilisant des effets dépendants de chaque modalité. Par exemple, d'un point de vue visuel, l'auteur peut jouer avec les couleurs, l'éclairage, l'angle, la distance et les mouvements de caméra. Mais il peut également jouer sur le volume, le rythme, et les styles musicaux de la bande son, ou encore jouer sur l'apparence, la couleur du texte. Finalement tous ces éléments de style permettent à l'auteur de faire passer son intention artistique.

#### 1.3.2.1 La description de la scène

La description de la scène correspond aux lieux, temps et actions de l'histoire. C'est en fait le décor ou concept du plan vidéo mais par extrapolation cela peut également être l'ambiance, l'atmosphère de ce segment vidéo ou même encore des actions particulières.

La modalité visuelle apporte beaucoup d'éléments pour extraire cette information de contexte. Le mouvement local, et donc l'analyse des séquences d'images, est moins déterminant pour cette tâche. En effet, le décor est un élément essentiellement statique alors que les personnages sont généralement des éléments dynamiques. C'est pourquoi les travaux sur la description de scène sont souvent issus de l'indexation des images statiques où les descripteurs de couleur et de texture sont largement utilisés (voir le §1.2.1). On trouve de nombreux travaux dans ce domaine. Par exemple dans [Szummer et Picard, 1998], ces descripteurs per-

mettent de différencier automatiquement les scènes d'intérieur des scènes d'extérieur. D'autres travaux comme ceux de [Vailaya *et al.*, 1998] s'intéressent aux scènes en extérieur et, grâce à l'utilisation d'histogrammes couleur et de mesures de cohérence sur les vecteurs de direction spatiale des contours, permettent de retrouver et discriminer les images de villes des images de paysages naturels. Ensuite, les mêmes auteurs dans [Vailaya *et al.*, 2001] ou d'autres comme [Szummer et Picard, 1998] ou [Snoek *et al.*, 2006] se focalisent sur les paysages naturels et retrouvent les *forêts*, les *montagnes*, la *mer*, les *plages*, le *désert*, les *chutes d'eau* mais également le *ciel* et les *couchers/levers de soleil*. Au delà des scènes naturelles ou urbaines, les travaux récents s'intéressent à la détection de concepts sémantiques qui correspondent habituellement à une description du cadre ou de l'action. Ces travaux, en particulier ceux développés dans le cadre du challenge TRECVID, s'intéressent à la détection de "**H**igh-**L**evel **F**eature" [Smeaton *et al.*, 2009] tel que *courir/marcher*, *fumer*, *boire*, des concepts comme la *violence physique*, les *catastrophes naturelles*, les *incendies*, etc. Ainsi plus de 101 concepts sémantiques sont recherchés automatiquement [Snoek *et al.*, 2006]. Voir [Smeulders *et al.*, 2000] et [Lavee *et al.*, 2009] pour un état de l'art des méthodes et descripteurs utilisés pour retrouver automatiquement des concepts sémantiques dans les vidéo.

En ce qui concerne la modalité audio, de nombreux travaux permettent de retrouver des environnements sonores particuliers. Par exemple dans [Zhang et Kuo, 1999], les auteurs sont capables de discriminer les sons naturels ou synthétiques par l'utilisation du timbre et du rythme. Ainsi la reconnaissance de sons spécifiques [Wold *et al.*, 1996, Lu *et al.*, 2002] comme la *pluie*, la *foule*, l'*eau* (rivière, mer, ...), le *tonnerre*, des *explosions*, permet ensuite de reconnaître des environnements plus globaux [Chu *et al.*, 2006] comme les *halls d'accueil* (ouverture/fermeture occasionnelle de portes, bruit éloigné des ascenseurs, individus parlant tranquillement), les *restaurants/cafés* (bruit de foule, sonnerie des caisses enregistreuses, déplacement des chaises), la *rue* (trafic des autobus et des voitures). De plus, la reconnaissance d'instruments de musique [Herrera-Boyer *et al.*, 2003] ainsi que la classification des genres musicaux [Tzanetakis et Cook, 2002] ont permis des travaux sur l'émotion ou l'atmosphère véhiculée par le son [Zentner *et al.*, 2008, Ruvolo *et al.*, 2008]. Pour [Petrushin, 1999, Lee et Narayanan, 2005] cette émotion est analysée à partir des dialogues entre les personnages.

La modalité texte peut également fournir des informations utiles pour la description de scène. Ces informations, très souvent des indications de lieu, de temps et/ou d'ambiance, peuvent être extraites des méta-données [Bulterman *et al.*, 2007, Buehler *et al.*, 2009]. Mais il est également possible d'obtenir des informations à partir du texte présent dans les images ou des paroles de la bande son. Des techniques spécifiques d'extraction de caractères ou d'analyse de la parole sont alors nécessaires pour obtenir ce texte [Snoek *et al.*, 2005, Bertini *et al.*, 2006]. Ce changement de modalité sera traité dans le §1.3.3.1

#### 1.3.2.2 La détection d'objets

Les objets dont nous parlons ici sont des entités statiques ou dynamiques dans le document vidéo. Comme dans le cas de la description d'une scène, les approches concernant la modalité visuelle sont souvent issues de la reconnaissance d'objets dans les images statiques mais utilisent également la détection locale de mouvements. Les descripteurs de

forme, de texture et de couleur sont très utilisés dans ces travaux pour retrouver des animaux comme les poissons [Mokhtarian *et al.*, 1997] ou de nombreux objets comme des bâtiments, voitures, bicyclettes, routes, arbres, etc. [Swets et Weng, 1996, Del Bimbo, 1999, Van de Sande *et al.*, 2008, Pujol, 2009]. Comme nous l'avons déjà évoqué, ces travaux font l'objet de challenges comme *"PASCAL Visual Object Challenge"* ou *"Mediamill Challenge"* ou encore dans *"TRECVID HLF"*.

Des objets spécifiques peuvent également être retrouvés en utilisant la modalité audio. Par exemple dans [Wold *et al.*, 1996] et [Zhang et Kuo, 1999] des tintements de cloche, les sonneries de téléphone, les aboiements de chien ou des pleurs peuvent être détectés. Mais le plus souvent, la modalité audio sert à reconnaître des événements comme répondre au téléphone, ouvrir une porte, reconnaître les pas d'une personne qui marche, des bruits de vaisselles, etc. [Istrate, 2003, Cristani *et al.*, 2007] ou dans le cas de vidéo sportives : une faute sifflée par l'arbitre, un service au tennis, un but au football, etc. [Xu *et al.*, 2008].

L'extraction d'information à partir des textes permet comme dans le cas de la description de la scène de retrouver des objets spécifiques. Cependant, la détection d'objets est limitée à un certain nombre d'objets bien spécifiques. En effet la tâche de détection de contenu est grandement simplifiée lorsque l'on sait ce que l'on recherche (par exemple, de la forêt ou des voitures) car les connaissances *a priori* permettent de se focaliser sur des attributs particuliers et pertinents. La tâche est donc plus difficile en l'absence de cette connaissance. D'ailleurs, un détecteur générique reste encore inaccessible et demeure le but ultime pour les chercheurs en analyse de documents vidéo.

#### 1.3.2.3 La détection de personnes

Cette partie peut être vue comme une sous partie de la détection d'objets. En effet les approches sont souvent assez proches de celles utilisées pour les autres détections de contenu. Beaucoup de travaux sur ce sujet sont issus des études faites en télésurveillance. Le principe de base consiste à détecter les visages ou les corps humains en se basant sur les connaissances *a priori* des formes, des textures et couleurs et sur l'entraînement d'algorithmes d'apprentissage [Zhao *et al.*, 2003]. Ces approches fonctionnent relativement bien et permettent de retrouver, comme dans [Snoek *et al.*, 2006], des personnages publics comme : G. Bush jr, Y. Arafat, J. Kerry, B. Clinton, etc. De même, les algorithmes de reconnaissance de la parole et d'extraction des dialogues entre les personnages à partir de la bande son sont maintenant au point et peuvent contribuer à la détection de personnes. La détection d'une personne dans un texte passe généralement par la recherche d'un nom propre dans le corpus textuel [Satoh *et al.*, 1999]. Pour être efficace, ces systèmes utilisent des techniques de traitement de la langue naturelle. Ces techniques se basent sur des dictionnaires, thésaurus, analyseurs lexicaux, syntaxiques et sémantiques. Nous reviendrons sur ces approches dans le chapitre consacré à l'analyse des textes et plus particulièrement dans la section consacrée à "l'extraction d'information à partir de texte" §4.4.

Nous venons de voir que la description de contenu permet, à l'heure actuelle, de contribuer à la détection et la description d'objets, lieux, personnages spécifiques. Les modalités visuelles et audio sont intéressantes pour détecter l'environnement dans lequel est situé le passage vi-


FIGURE 1.4 – Image de Times Square issue d'un document vidéo

déo. L'utilisation de la modalité textuelle permet, quant à elle, de décrire plus précisément les éléments mis en œuvre (lieu géographique, nom des personnes, ...). L'utilisation conjointe (ou fusion) de ces informations issues des différentes modalités est donc nécessaire pour obtenir une description pertinente, riche et sémantique du document cinématographique. Par exemple sur la figure 1.4, l'image apporte l'information de scène (ville), car il est possible de détecter des immeubles, des voitures et des personnes. L'audio peut confirmer cela en précisant que c'est une ville américaine (bruit de sirène d'ambulance par exemple). Cependant, bien qu'une personne reconnaîtrait immédiatement que la ville est New York (taxis jaunes, buildings) et une vue de Times Square (panneaux lumineux), l'analyse automatique à partir de ces deux modalités n'apporte pas cette précision. Seule l'utilisation du texte dans ce cas apporte cette information (lieu et temps de la scène précisés dans l'image). Nous voyons à travers cet exemple que la fusion de ces différentes modalités peut fournir des informations caractérisant le document vidéo avec un plus haut niveau sémantique que par l'utilisation d'une seule de ces modalités. Cette analyse multimodale va être présentée dans la prochaine section.

### 1.3.3 L'analyse multimodale

Après la segmentation du document vidéo dont le résultat est la structure du film en plans vidéo puis l'extraction et la description du contenu de chacun de ces segments, nous avons vu l'intérêt d'utiliser conjointement les différentes modalités afin d'obtenir une meilleure caractérisation du document. La problématique maintenant est d'arriver à fusionner ces différentes informations pour arriver à une description sémantique du film. Cependant, les informations extraites des différentes modalités ne sont pas toujours de même niveau sémantique, et il est souvent nécessaire, dans une étape préalable, de convertir ces informations avant de les fusionner.

#### 1.3.3.1 La conversion de modalité

Afin d'extraire le contenu d'une modalité il est parfois préférable de convertir cette modalité. Cette conversion est généralement faite vers la modalité textuelle. En effet, nous avons vu dans la section précédente que cette modalité apportait bien souvent une information d'un niveau sémantique supérieur aux autres modalités.

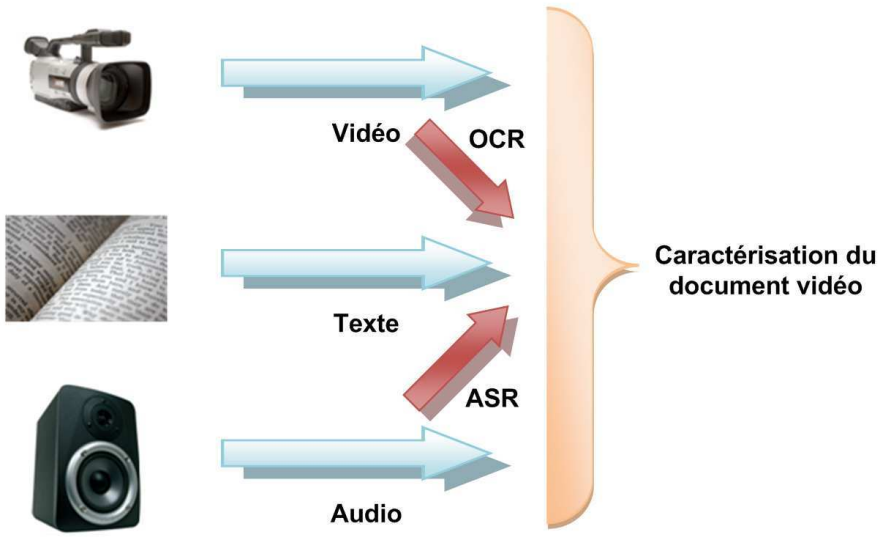



FIGURE 1.5 – Conversion de modalité et fusion des informations de contenu

Classiquement, la conversion du texte présent dans les images au texte passe par l'utilisation d'algorithmes de **reconnaissance optique de caractères** ou Optical Character Recognition (OCR) (voir figure 1.5). Il faut préalablement détecter la présence de texte dans les images puis localiser ce texte. Il existe principalement deux techniques basées sur la détection de régions ou sur la détection de textures. La détection de régions est basée sur les caractéristiques colorimétriques (les couleurs du texte en opposition avec la couleur du fond). Dans [Lienhart et Wernicke, 2002] les lignes de texte sont détectées en exploitant les différences de contraste. La détection de texte basée sur la texture, utilise la différence de texture entre le texte et le fond de l'image [Zhong *et al.*, 1995] (pour un état de l'art sur l'extraction et l'utilisation de texte à partir des images voir [Jung *et al.*, 2004, Yan et Hauptmann, 2007]).

La modalité textuelle est également obtenue à partir de la bande son lorsque celle ci contient un dialogue. Après une détection de la parole dans le support audio, les différents phonèmes sont extraits afin de reconstruire les mots. Ces systèmes, baptisés **reconnaissance automatique de la parole** ou Automatic Speech Recognition (ASR), donnent de bons résultats (de $\simeq 90\%$ de taux de reconnaissance pour l'anglais [Xie *et al.*, 2004] et $\simeq 85\%$ pour le finnois à $\simeq 40\%$ pour l'arabe Égyptien [Creutz *et al.*, 2007]) dans des conditions idéales, sans bruits ou perturbations. Mais les résultats sont moins bons (30% de taux d'erreur) lorsque le fond sonore est bruité (musique, foule, etc.) [Hauptmann *et al.*, 2002].

#### 1.3.3.2 La multimodalité

Le principe de l'analyse ou de l'intégration multimodale est d'améliorer la caractérisation du document par l'apport d'informations redondantes, complémentaires ou nouvelles. Ainsi l'ajout d'une modalité peut se décliner suivant trois buts :

- la **vérification** des informations. Lorsque les informations sont redondantes cela permet à partir de deux modalités de vérifier les informations issues de la troisième modalité.

- la **compensation** des informations. Lorsque les informations sont complémentaires cela permet par exemple de préciser et de compenser l'imprécision d'une information à partir des deux autres modalités.

- l'**ajout** d'information. Lorsque les informations sont différentes cela permet de caractériser plus largement le document et peut permettre par exemple, de faire du raisonnement à partir de bases de connaissances. Prenons l'exemple d'une vidéo personnelle dont le titre (méta-donnée) est : *"Vacances 2008 aux Etats Unis"*. Une image extraite de cette vidéo pourrait être la figure 1.4. Une analyse et une reconnaissance des objets sur cette image conduirait par exemple à trouver la voiture au premier plan dont la caractéristique couleur est jaune. Les informations extraites des autres modalités pourraient être : une information sur la scène (dans une rue en ville) à partir de l'audio (klaxons, bruits urbains), le pays (aux USA) à partir des méta-données et enfin la ville (à New York) où se déroule la scène à partir du texte incrusté dans l'image. Finalement, pour caractériser l'objet détecté dans l'image il serait possible de faire un raisonnement simple à partir d'une base de connaissances comme suit :

  $$USA + NewYork + Rue + Voiture + de\ couleur\ jaune \Longrightarrow Taxi$$

  Par cet ajout d'information multimodale le système pourrait caractériser automatiquement la voiture détectée dans l'image comme étant un taxi New Yorkais (Yellow Car).

Pour atteindre l'intégration et la fusion de ces informations, [Snoek et Worring, 2005] proposent de classer les différentes approches selon trois catégories distinctes :

- **L'extraction de contenu** peut être **symétrique** ou **asymétrique**. Lorsque les extractions de contenu de chacune des modalités sont indépendantes, elles sont dites symétriques. Les informations sont extraites parallèlement et les étapes d'extraction ne sont pas mises en cascade comme dans le cas asymétrique. Dans ce dernier cas, les extracteurs sont mis en série et il y a donc interaction entre eux. L'information extraite d'une modalité va orienter l'extraction d'information de(s) autre(s) modalité(s).

- **Le processus d'intégration** peut être **itératif** ou **non itératif**. Lorsque l'intégration des informations se fait par cycles, le processus d'intégration est dit itératif, comme dans [Naphide et Huang, 2001] où les informations et leurs interactions sont modélisées par des poids dans un réseau bayésien modifiés itérativement. Notons que les approches basées sur des processus d'intégration itératifs sont plus rares.

- **La méthode d'intégration** (on parle aussi de fusion de données) peut être basée sur des **connaissances** *a priori* où l'intégration des données est obtenue par une expertise

et des règles de combinaison. Dans [Tsekeridou et Pitas, 1999] la combinaison entre les informations audio et vidéo (détection de la parole, de silence, de l'identité de la personne qui parle, présence d'un visage, absence d'un visage, présence d'un visage qui parle) apporte à l'utilisateur des informations beaucoup plus détaillées. Par exemple la personne X parle et est présente dans la scène Y (caractérisée par sa durée) ou un reportage apparaît dans la scène Z (caractérisée par sa durée) tandis que le journaliste W raconte l'histoire sans être présent dans la scène. Les informations audio et visuelle contiennent des contenus différents qu'il est possible de combiner grâce à des règles d'interaction comme celles de l'exemple ci dessus. Dans [Valet, 2001] la détection de zones d'intérêt dans des images sismique est obtenue par raisonnement flou imitant celui des experts du domaine. Les travaux utilisant cette méthode d'intégration sont plus rares et les descripteurs issus des extracteurs d'information doivent être explicites et compréhensibles par l'homme afin de leur appliquer des règles de fusion issues d'une connaissance du domaine. L'intégration des informations est obtenue également par l'utilisation de **classifieur** lorsque de telles connaissances ne sont pas disponibles. De nombreux algorithmes de classification ont été utilisés (SVM, réseau de neurones, réseau bayésien ...). Dans ces approches chaque concept est caractérisé par un ensemble de descripteurs vus comme les axes d'un hyper-espace. Le but est de trouver les régions permettant de séparer et caractériser les concepts suivant ces descripteurs.

Un bon aperçu des travaux couvrant ces différentes approches est présenté par l'état de l'art de [Snoek et Worring, 2005]. Notons tout de même que la grande majorité des travaux d'analyse multimodale sont du type symétrique avec des méthodes de classification statistiques dans des processus de fusion non itératifs. De plus, la synchronisation et l'alignement des modalités sont fondamentaux pour permettre une intégration et une fusion des informations extraites.

## 1.4 Conclusion

Pour conclure ce chapitre, l'utilisation conjointe des différentes modalités audio, image et texte est nécessaire pour caractériser efficacement les documents vidéo et plus précisément les films. Cette caractérisation passe par l'étape de segmentation permettant de retrouver les segments vidéo dont le contenu est sémantiquement homogène. La description de ces plans video est basée sur la détection d'éléments spécifiques dont la caractérisation et la description sémantique se font le plus souvent par l'intermédiaire de la modalité textuelle (généralement non disponible avec le film sauf après conversion de modalité). De plus, la connaissance et l'apport d'information *a priori* du domaine permet d'améliorer les processus d'extraction et de fusion d'information et permet in fine une caractérisation plus précise du film. Cependant, la majorité des travaux présentés précédemment ont été testés et développés sur des données applicatives constituées de films "classiques", c'est-à-dire mettant en scène des personnes (êtres humains) dans des environnement réels. Nos travaux se démarquent des travaux précédents par la caractérisation de films d'animation dont les caractéristiques se distinguent des films "classiques" sur de nombreux aspects. C'est ce que nous allons voir dans le chapitre suivant.

Chapitre 2

# Le contexte de travail : les films d'animation

**Résumé :** Dans ce chapitre nous présentons le contexte de travail de ces travaux de thèse liés au domaine applicatif du cinéma d'animation. Nous abordons les problématiques d'indexation et de caractérisation de ces films particuliers que sont les séquences d'animation. Nous présentons les caractéristiques de ces films ainsi que celles de la base de données vidéo et textuelle dont nous disposons. Puis nous présentons les problématiques et les solutions apportées dans nos travaux.

## 2.1 Présentation du contexte de travail

Nous venons de voir les défis que posent l'indexation de documents multimédia et les solutions mises en œuvre, en particulier dans le cas de films. C'est dans ce contexte que se situent les travaux de cette thèse, avec comme spécificité leur application à la base numérisée de films d'animation mise en place par la Cité de l'Image en Mouvement (CITIA) dans le cadre du Festival International du Film d'Animation d'Annecy. Aussi, avant toute chose, nous allons voir quelles sont les caractéristiques de ces documents vidéo particuliers.

### 2.1.1 CITIA et la base de films d'animation

Portée par la communauté de l'agglomération d'Annecy, le département de la Haute-Savoie et la Région Rhône Alpes, CITIA [CITIA, 2009b] tire ses origines du Festival International du Film d'Animation d'Annecy (FIFA) qui confère à Annecy depuis plus de 45 ans une renommée mondiale dans le domaine du film d'animation. Le Centre International du Cinéma d'Animation (CICA), association Loi de 1901 créée en 1984, a constitué les fondements de ce projet à travers ses différentes missions : organisation du festival et du Marché International du Film d'Animation (MIFA), promotion, diffusion et soutien du cinéma image par image, développement et exploitation d'un centre de documentation multimédia.

Dans le contexte de cette dernière mission, CITIA, Établissement Public de Coopération Culturelle, regroupe sous forme numérisée l'important fond documentaire du CICA. La

constitution de cette base numérique des films d'animation est en cours de constitution (numérisation et stockage des films) et contiendra à terme 30000 films auxquels s'ajouteront annuellement les quelques centaines de films mis en compétition lors de chaque festival. De cette façon les professionnels de l'animation et les écoles spécialisées pourront bientôt avoir accès à ce fond et notamment à des extraits de films d'animation via internet. L'exploitation de cette base et l'élaboration de méthodes documentaires associant base de connaissances textuelles et analyse automatique des films d'animation constituent le cadre d'un partenariat entre le Laboratoire d'Informatique Systèmes, Traitement de l'Information et de la Connaissance (LISTIC) et CITIA [CITIA, 2009c]. Ce contexte pluridisciplinaire et cette base particulière contribuent à l'originalité de nos travaux.

L'indexation et la caractérisation des documents numériques que sont les films d'animation du CICA, constituent le cadre méthodologique et applicatif des travaux présentés dans ce manuscrit. Nous avons vu précédemment que l'utilisation d'information *a priori* et de connaissances sur les documents à indexer améliorent la caractérisation automatique des vidéos. Nous allons donc préciser dans les sections suivantes quelles sont les caractéristiques particulières des films d'animation de CITIA et en quoi ces films diffèrent des films classiques.

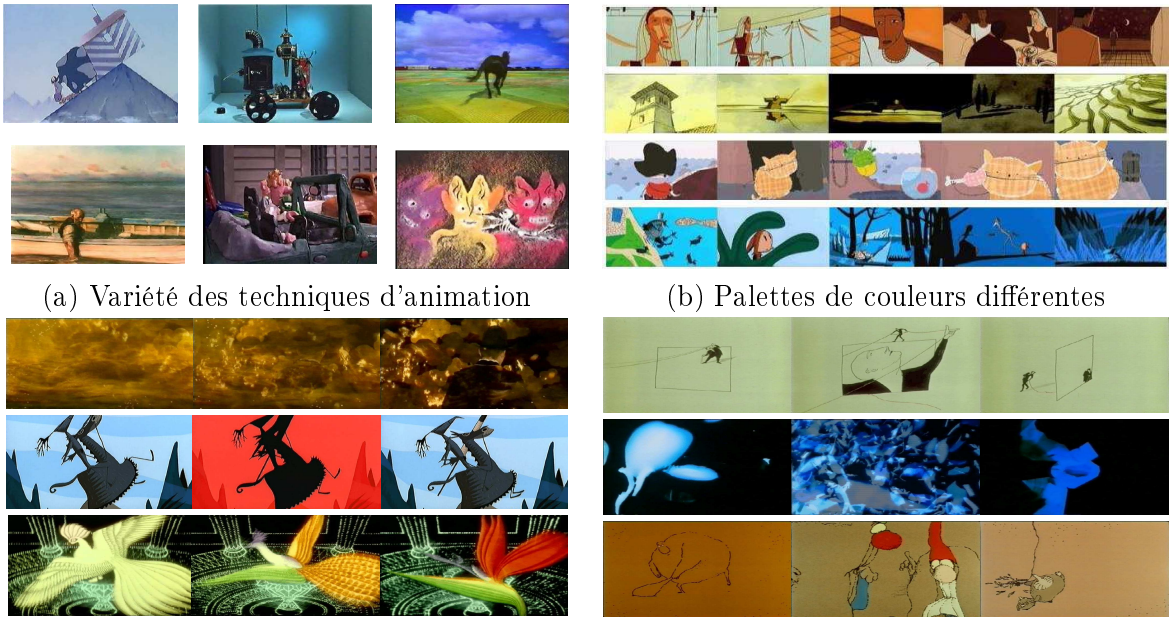
### 2.1.2 Les films d'animation

Le cinéma est un art qui offre au public une œuvre (ou film par métonymie) composée d'images en mouvement généralement projetées à la cadence de 24 images par seconde. C'est la succession rapide d'images différant en moyenne peu les unes des autres qui fournit au spectateur l'illusion d'une image animée, reproduisant les mouvements et trajectoires de la vie réelle. Grâce à la persistance rétinienne et à l'effet phi, les techniques de projection permettent à l'être humain de voir cette série d'images discrètes comme un flux visuel continu. De ce principe de base est né l'animation ou cinéma d'animation. Cet art regroupe toutes les œuvres (ou films) dans lesquelles l'auteur donne l'illusion de la vie à des objets qui par nature sont inertes. Nous voyons donc apparaître une différence significative avec les films que l'on nommera « classiques » car mettant en scène principalement des êtres humains et/ou animaux (bien que les films d'animation puissent aussi utiliser des images naturelles). En partant de ce principe, les facettes de l'animation sont pratiquement sans limite. Du point de vue artistique l'auteur propose, par l'intermédiaire de son œuvre, la communication d'une émotion au spectateur qui regarde son film d'animation. Cette intention artistique, dans le cas des films d'animation, s'exprime de différentes manières, depuis l'utilisation de techniques d'animation particulières jusqu'aux genres et sujets traités. Dans ce qui suit, parmi ces différentes formes d'expression, nous nous intéresserons plus particulièrement à celles qui intéressent la mise en place des techniques d'indexation automatiques.

#### 2.1.2.1 Les techniques d'animation

Depuis l'invention du cinéma d'animation à la fin du XIX$^{e}$ siècle, les artistes n'ont pas manqué d'imagination pour développer des techniques leur permettant de créer l'illusion du mouvement image par image. Depuis le dessin sur papier, en passant par la pâte à modeler, pour arriver à la pixilation puis à l'image de synthèse, tout est permis et aujourd'hui encore de nouvelles techniques apparaissent (l'*annexe B.5* donne une répartition des films de la base CITIA en fonction des techniques d'animation). Ces techniques, souvent très dif-

férentes, rendent très difficiles les analyses de contenu à partir des séquences d'images. En effet, les textures, couleurs, contrastes, etc, en un mot les caractéristiques image sont fortement influencées par les techniques d'animation (voir figure 2.1.a) et les effets spéciaux qui en découlent (voir figure 2.1.c). C'est l'utilisation de ces techniques particulières qui distingue fondamentalement le cinéma d'animation du cinéma classique. Le détail en images de ces techniques est disponible en *annexe A*.



(a) Variété des techniques d'animation

(b) Palettes de couleurs différentes

(c) Effets visuels

(d) Complexité du contenu

FIGURE 2.1 – Les particularités des films d'animation (source : les films de CITIA).

#### 2.1.2.2 Les couleurs

L'usage de la couleur fait partie de l'éventail des possibilités que les auteurs ont à leur disposition pour faire passer leurs intentions artistiques. Dans les films d'animation, il est assez fréquent de trouver l'utilisation d'une palette de couleurs particulière contenant un nombre réduit de couleurs. C'est une des différences majeures avec les images/vues réelles. En effet il est techniquement très difficile pour l'artiste de peindre la totalité des variations colorées que l'on retrouve dans une image réelle (voir figure 2.1.b). Bien que l'arrivée des images numériques ait bouleversé les techniques de colorisation (dans une image numérique, le choix des couleurs se fait parmi plus de 16 millions de couleurs), l'artiste utilise généralement une palette de couleurs réduite pour construire sa séquence animée. Souvent, le choix et la distribution de ces couleurs traduit l'intention artistique de l'auteur. Cela permet de transmettre certains sentiments ou sensations comme la chaleur, l'harmonie, le contraste, la joie, la tristesse, etc. On retrouve de nombreux travaux sur la théorie des couleurs depuis l'Antiquité jusqu'à nos jours en passant par Newton ou Goethe. Par exemple, le physicien Rumford fut, dès 1797, le premier à affirmer que les couleurs n'étaient harmonieuses que si leur mélange donnait du blanc. Plus tard dans "Le cercle chromatique d'Itten" [Itten, 1974], Johannes Itten tente de rationaliser l'utilisation de la couleur et ses contrastes chez les artistes peintres. Dans sa théorie, le contraste de la couleur peut aussi bien exprimer une joie

débordante qu'une profonde tristesse. À angle droit avec l'axe jaune / violet, sur le cercle chromatique (voir figure 2.2), se trouvent les couleurs rouge-orange (couleur la plus chaude) et bleu turquoise (couleur la plus froide). Le contraste chaud-froid le plus fort est obtenu en juxtaposant ces deux couleurs. Dans le même ordre d'idées, il a été montré que dans une pièce peinte en bleu-vert, les personnes trouveront qu'il fait froid à 15°C alors que dans une pièce rouge-orangé, elles ne ressentiront le froid qu'à 11-12°C. Pour Birren [Birren, 1969] il y a une loi de l'harmonie des couleurs. Selon lui, la beauté résulte d'un bon ordonnancement des couleurs.

Dans les films d'animation, l'utilisation d'une palette de couleur réduite permet d'utiliser des techniques d'analyse mettant en évidence les couleurs dominantes, les contrastes, les harmonies, etc. ([Ionescu, 2007])

FIGURE 2.2 – Roue chromatique de Johannes Itten

#### 2.1.2.3 Les genres

Une autre des caractéristiques singulières des films de la base s'exprime à travers les genres traités. En effet, les genres sont nombreux (ces films ne sont pas tous axés sur le divertissement) et les frontières entre les genres pas toujours très nettes, créant ainsi une multitude de contenus difficiles à analyser automatiquement. Bien que l'on retrouve en partie la diversité des sujets traités dans les films classiques (genre policier, humoristique, comédie, aventure, etc.), il prédomine néanmoins dans les films de CITIA une volonté artistique qui aboutit souvent à des films très originaux dont le contenu n'est pas toujours facile à classer dans une catégorie. De plus, le fait de donner vie à des objets inertes par nature, permet une création sans limite puisque l'artiste ne subit pas les contraintes naturelles de notre monde (pesanteur, continuum espace-temps, etc.) rendant sans fondement un certain nombre d'hypothèses de mouvement par exemple. Les spécialistes de l'animation estiment que 30% des films d'animation présentés au festival ne peuvent pas être résumés tant leur contenu est singulier (voir figure 2.1.d) (information recueillie durant une conversation privée auprès de GIANNALBERTO BENDAZZI professeur à l'université de Milan spécialiste du cinéma d'animation [1]). Cette particularité des films d'animation rend donc quasi impossible leur analyse automatique sans information externe. Dans l'annexe B.6, nous donnons la répartition des films dans la base selon le genre déclaré.

1. http://www.lapisvillage.net/static/cur_bendazzi.htm

#### 2.1.2.4 Bilan

On l'a vu, les films de CITIA sont différents des films naturels et des films "grand public" d'animation (communément appelés dessins animés). Cette différence tient essentiellement à leur contenu qui relève souvent d'une intention artistique plus que d'une recherche de divertissement. Les caractéristiques les plus importantes des films d'animation peuvent se résumer de la manière suivante :

- **Les techniques d'animation :** sont spécifiques et peuvent être mixées entre elles (voir figure 2.1.a et annexe B.5).
- **Les couleurs :** marquent une volonté de l'auteur et sont une signature de l'œuvre et/ou de l'artiste (voir figure 2.1.b).
- **Le contenu :** présente une variété extrême (voir figure 2.1.d et annexe B.4).
- **Les événements :** ils ne suivent pas forcément une chronologie bien établie (continuum espace-temps). Des objets peuvent apparaître ou disparaître de la scène, se mettre à léviter, et les personnages peuvent courir dans les airs. Tout est possible et ne dépend que de l'imagination de l'artiste.
- **Les personnages :** si il y en a, ils peuvent prendre n'importe quelle forme, couleur ou texture.
- **Les effets spéciaux :** certains effets sont propres au cinéma d'animation comme par exemple le "Short Color Change (SCC)", brusque variation de couleur dans un même plan (voir figure 2.1.c).
- **La durée :** les films d'animation sont généralement du type court métrage d'une durée moyenne de 10 minutes (voir annexe B.2).

Toutes ces caractéristiques, en l'absence d'information externe supplémentaire, rendent très difficiles la réalisation des tâches d'indexation et de caractérisation automatique des films d'animation. L'extraction de descripteurs de haut niveau sémantique à partir de l'image (comme la détection de visage) est très délicate, compte tenu de la grande variabilité des caractéristiques image et du manque d'information *a priori*. En effet, les informations *a priori* de forme, texture, couleur qui permettent à de nombreux détecteurs de retrouver des visages ou autres objets spécifiques comme des voitures, des buildings, etc, ne fonctionnent plus sur la majorité des films d'animation car ces connaissances issues du monde réel, ne sont pas toujours valides dans le monde de l'animation. C'est ce que nous avons baptisé le *"paradoxe de Donald"*. Nous voyons dans les images de la figure 2.3 que les caractéristiques de forme, de textures, et de couleurs que l'on pourrait apprendre de la figure 2.3.a ne sont plus des attributs image pertinents pour retrouver un canard sur les images issus de films d'animation (figure 2.3.b à .f).

Dans la base des films de CITIA, le stockage d'un film est complété par sa fiche d'inscription, dans laquelle est rassemblé un certain nombre d'informations directement liées au film. Cette ressource textuelle, dont la présentation détaillée est faite dans la section suivante, peut

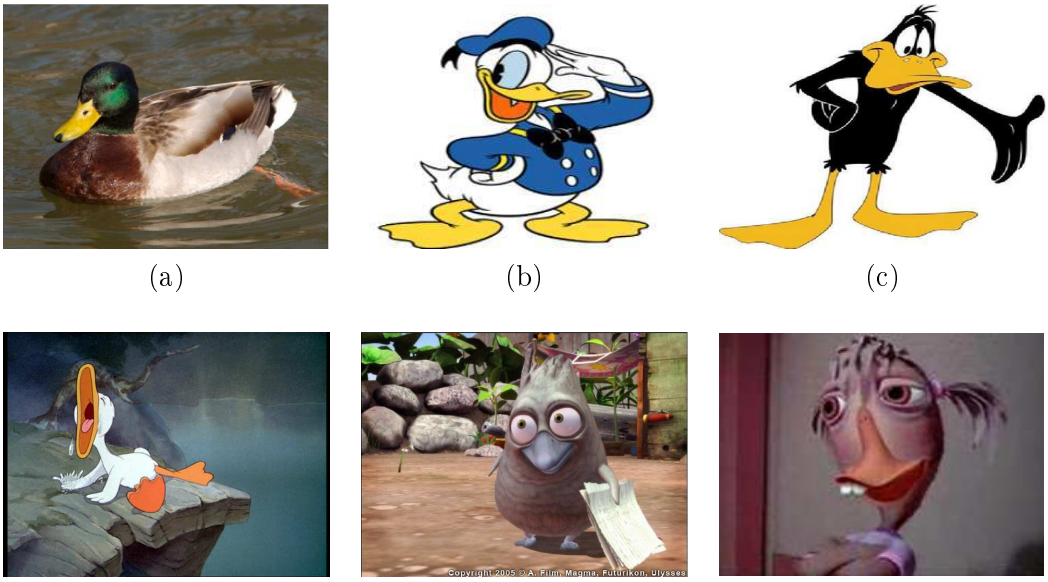



FIGURE 2.3 – Différents canards : du réel à l'animation. *(a) : Canard colvert, (b) : Donald Duck de Disney (1934), (c) Daffy Duck de Tex Avery (1937), (d) : Le vilain petit canard de Disney (1931), (e) : Le vilain petit canard et moi de Gebeka Films (2005), (f) : Chicken Little de Disney (2004)*

être très utile pour la caractérisation des films. Dans nos travaux, nous proposons d'exploiter l'information issue de ces péritextes, en complément des informations extraites des films, afin d'apporter une information de haut niveau sémantique.

### 2.1.3 Les fiches d'inscription

Pour inscrire un film à la sélection du festival, les auteurs doivent fournir une fiche d'inscription contenant un certain nombre d'informations concernant le film. L'ensemble de ces fiches est regroupé pour former une base textuelle accessible depuis un moteur de recherche, appelé Animaquid, disponible sur le site de CITIA ([CITIA, 2009a]). On retrouve dans ces fiches des informations essentielles comme :

- Le titre du film (dans la langue originale, en français et en anglais), la nationalité et l'année de production.
- Les noms des auteurs : scénario, graphismes, sons, etc.
- Des indications sur les techniques utilisées, l'âge du public visé, le genre du film, sa durée, son support.
- Un synopsis en français et en anglais qui est un court texte descriptif du sujet traité par le film (en moyenne 20 mots voir annexe B.7).

La figure ci-dessous (figure 2.4) présente, à titre d'exemple, la fiche du film "Au bout du monde".

**Titre** *Au bout du monde (At the End of the Earth)*

**Synopsis** *Posée sur le pic d'une colline, une maison balance alternativement de droite à gauche, au grand dam de ses habitants. To the great displeasure if its inhabitants, a house set on top of a hill sways from left to right.*

**Identité** *Réalisation : Konstantin BRONZIT, Pays : France, Année : 1998, Durée : 07 mn 45 s*

**Technique** *Technique(s) utilisée(s) : Dessin sur cellulos, Procédé : Couleur, Version : Sans dialogue ni commentaire, Catégorie : 1999 Courts métrages, Genre(s) : Humour, Public(s) visé(s) : Tout public*

FIGURE 2.4 – Fiche d'inscription au FIFA du film *"Au bout du monde"*

Malheureusement ces fiches présentent quelques imperfections :

- les données sont parfois incomplètes, en particulier pour les films les plus anciens : des champs - genre, technique, synopsis - ne sont pas renseignés).

- Les indications ont été le plus souvent remplies par ceux qui ont inscrit le film au festival, mais aussi parfois par les personnes qui ont saisi les données dans la base, ce qui atténue la fiabilité de certaines de ces informations, en particulier quand il s'agit du genre ou du résumé.

- Les synopsis (qui sont des textes courts) présentent une très grande variabilité dans la manière dont ils sont rédigés. Selon les cas, au lieu d'être un résumé, cela peut être une accroche voire dans certains cas une pensée philosophique (*« To be eaten or not to be eaten ? That is the question ! »* pour le film *Circuit Marine*).

- La qualité des traductions est variable selon les périodes de saisie.

- Les techniques et les genres renseignés sont quelquefois insolites. Par exemple les genres *personnel*, *afrique*, *spirituel*, etc., ne sont pas (et ne peuvent pas être) référencés dans l'ontologie des genres[2] [Beauchêne et Deloule, 2009].

En conclusion, nous avons donc un corpus dont la qualité n'est pas homogène. Néanmoins, cette ressource textuelle est très intéressante à exploiter pour caractériser les films car elle permet d'apporter des informations *a priori*, de haut niveau sémantique, sur le document vidéo, informations difficilement disponibles par ailleurs tant les caractéristiques des films sont particulières. Le contexte général étant présenté nous allons préciser maintenant les objectifs de ces travaux.

2. Une ontologie des genres a pu être construite par l'équipe Condillac du LISTIC avec l'aide des experts de l'animation

## 2.2 Présentation des objectifs

Afin d'exploiter efficacement cette base de films d'animation, il est nécessaire de disposer d'outils logiciels plus performants que l'outil (Animaquid) actuellement utilisé. En effet, Animaquid n'utilise que des éléments textuels (Titre, Auteur, etc.) issus des bulletins d'inscription au festival. Notre travail est une participation au développement de ces nouveaux outils logiciels. Plus spécifiquement, les travaux effectués dans cette thèse ont pour objectif la caractérisation des films d'animation. Cette caractérisation, dont la qualité détermine l'efficacité des outils d'exploitation de la base, est construite à partir de la fusion d'informations liées aux films.

Les informations que nous avons exploitées sont extraites de trois sources :

- **les séquences d'images** : des descripteurs de type couleur, texture, forme, mouvement, etc. peuvent être extraits de l'analyse des images. Ces descripteurs restent d'un niveau sémantique relativement faible. En effet, comme nous l'avons déjà évoqué, la recherche de descripteurs de plus haut niveau, comme des visages humains, des voitures, etc. est difficilement envisageable dans le cas des films d'animation (c'est le "paradoxe de Donald").

- **les péri-textes** : des informations sur le film et son contenu peuvent être extraites des textes issus des fiches d'inscription. En particulier, les synopsis permettent d'accéder à des informations dont le niveau sémantique est élevé.

- **L'expertise du domaine** : en effet la connaissance des experts du cinéma d'animation est une information importante dans le processus de fusion d'information.

L'objectif principal (voir figure 2.5) et l'originalité de cette thèse est la mise en place de méthodologies permettant de **fusionner** les informations extraites de ces deux modalités.

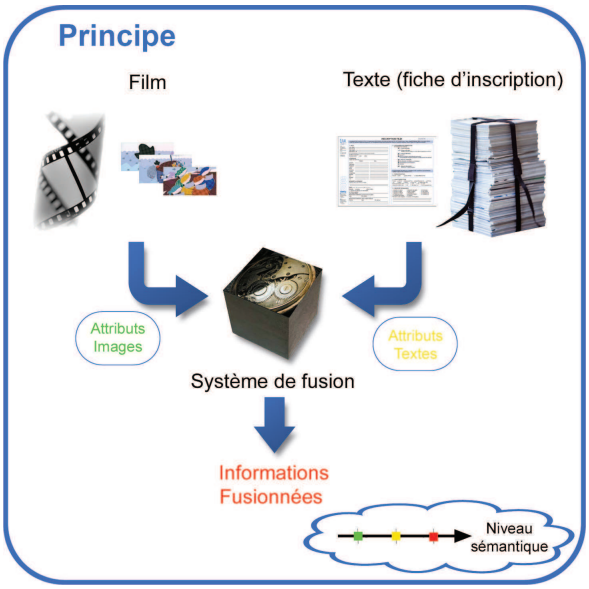



FIGURE 2.5 – Principe de la fusion d'information texte et image

A notre connaissance, peu de travaux en indexation vidéo ont tenté ce type de fusion multimodale. Cette fusion pose en effet un certain nombre de difficultés :

- **Une différence de niveau sémantique** : l'analyse des images apporte une information dont le niveau sémantique reste faible alors que les synopsis contiennent au contraire une information dont le contenu sémantique est élevé.

- **Une désynchronisation** : en effet, les informations obtenues à partir des synopsis ont généralement un lien sémantique avec les images mais ce lien n'est pas repéré sur l'échelle temporelle de la vidéo (voir §1.3.3.2). Ainsi, un synopsis pourra parler du "début de l'histoire", ce qui reste très imprécis comparé à la précision de synchronisation que l'on a par exemple sur des sous-titres.

- **Une différence dans la quantité d'information** : les images apportent une très grande quantité d'information (plusieurs milliers d'images par séquence vidéo) alors que les synopsis contiennent au contraire une information en quantité très réduite ($\simeq$ 20 mots).

Parmi les sources d'information disponibles, il y a également la bande son du film et, éventuellement, le texte obtenu après conversion de modalité (§1.3.3.1). Nous avons choisi de ne pas exploiter ces sources. En effet, les films d'animation ne contiennent pas toujours de dialogues et ces derniers peuvent être dans n'importe quelle langue (langue d'origine du pays de production). De plus, les musiques ou bruit d'ambiance présentent une grande variabilité qui les rendent difficilement exploitables. Quant aux textes, ils sont très rares et leur exploitation ne se justifie donc pas.

Ce mémoire s'articule autour de deux parties :

**L'extraction d'information** est abordée à travers :

- l'analyse des séquences d'images dans le **chapitre 3**, où des informations de couleur et d'activité sont extraites de la modalité vidéo.

- l'analyse des textes dans le **chapitre 4**, où des informations d'atmosphère et de description du film sont extraites des synopsis.

**La fusion d'information** est abordée dans le **chapitre 5** à travers la fusion des informations issues de l'analyse des images et de l'analyse des textes afin d'obtenir une caractérisation sémantique des films. Cette caractérisation des films est faite à un niveau global où nous cherchons à caractériser l'atmosphère dégagée par le film et à un niveau local où nous cherchons à décrire grâce au texte les passages d'action.

# Deuxième partie

# Extraction d'information

Chapitre 3

# Extraction d'information à partir des images

**Résumé :** Dans ce chapitre consacré aux images nous abordons la caractérisation des séquences d'animation par extraction d'informations sur la couleur (caractéristique importante des films d'animation) et sur l'activité dans ces séquences d'images. La première partie de ce chapitre fait la synthèse des applications images transposables à notre contexte d'étude puis une seconde partie détaille notre approche et son fonctionnement. Notre solution est basée sur un algorithme dit à accumulation d'erreur qui permet d'extraire d'une séquence d'images un ensemble d'images clefs permettant à travers différentes étapes et traitements, la mesure des caractéristiques recherchées.

Dans les travaux présentés dans ce manuscrit, la tâche d'indexation des films d'animation passe par l'extraction d'information et la caractérisation du document à partir des modalités image et texte. Dans ce chapitre nous nous intéresserons à la caractérisation des séquences d'animation à partir des images. Cette caractérisation passe par l'analyse et l'exploitation des couleurs dans les images ainsi que par la recherche d'une caractérisation de l'activité dans les séquences vidéo.

## 3.1 L'existant

Nous avons déjà vu dans le chapitre 1.2.1 un certain nombre d'approches permettant d'indexer les images ou les séquences d'images. Ces approches sont étroitement liées à la nature des images traitées (image naturelle, vue extérieure, vue intérieure, paysage, ville, etc.) et s'appuient le plus souvent sur des connaissances *a priori* de couleur, texture, forme, mouvement, etc. (par exemple : la pelouse est généralement verte, les villes sont constituées d'immeubles ce qui entraîne la présence de lignes verticales dans les images). Dans le cas particulier des films d'animation ces connaissances *a priori* sont fortement remises en cause. Ainsi nous allons dans un premier temps détailler les approches classiquement utilisées dans la caractérisation d'images et discuter de leur application possible dans le cas des films d'animation.

### 3.1.1 Les grandes approches et leurs possibles applications aux films d'animation

Nous avons vu dans le chapitre sur l'indexation que l'extraction d'information à partir des séquences d'images peut être décomposée en deux parties. La première consiste à extraire l'information à partir des images. La seconde consiste quant à elle à utiliser la dimension temporelle des vidéos (les séquence d'images). Dans le cas de l'extraction d'information à partir des images un certain nombre de descripteurs peuvent être extraits automatiquement à partir des propriétés suivantes :

**Les couleurs** : elles jouent un rôle très important dans la transmission d'informations visuelles. En effet, l'œil humain est plus sensible aux changements de teinte des couleurs qu'à la présence de mouvement. De plus, les films d'animation constituent un type particulier d'expression artistique. Chaque film a sa propre distribution des couleurs voulue par l'auteur (voir 2.1.2.2). Dans un film d'animation, l'artiste choisit les couleurs qu'il va utiliser pour composer son œuvre en concordance avec son projet artistique. Ainsi, les couleurs prédominantes utilisées dans la séquence, la combinaison de ces couleurs, les impressions transmises, etc. sont des caractéristiques intéressantes à exploiter [Ionescu *et al.*, 2005a]. L'analyse des couleurs dans les séquences d'images est donc une orientation privilégiée dans notre contexte applicatif.

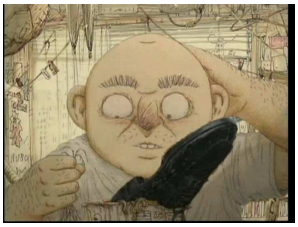
**Les formes** : elles sont décrites par leurs propriétés géométriques globales ou structurelles et les méthodes d'analyse utilisées sont généralement basées sur des approches contours ou des approches régions [Zhang et Lu, 2004]. Bien que les méthodes d'extraction de formes soient envisageables dans notre contexte applicatif, la difficulté majeure réside dans le passage de ces descriptions géométriques à des concepts exploitables. En effet, les connaissances *a priori* et les propriétés géométriques bien connues de certains objets de notre monde (par exemple une voiture, un bâtiment, un arbre ou un visage) permettent par des méthodes d'apprentissage de retrouver ces formes et la reconnaissance des concepts dans les images issues des films plus conventionnels [Mokhtarian *et al.*, 1997]. La détection automatique de visage par exemple utilise très souvent ces informations de formes [Zhao *et al.*, 2003]. Malheureusement dans le domaine de l'animation, ces approches sont confrontées à l'extrême variabilité des caractéristiques de formes des objets recherchés. Par exemple la détection de visage de personnages d'animation à partir des caractéristiques de forme semble beaucoup plus complexe à mettre en œuvre lorsque l'on regarde la diversité des formes utilisées dans le panel de visages présenté sur la figure 3.1. Ainsi, à cause du *"paradoxe de Donald"* il est peu probable que cette caractéristique soit intéressante dans l'immédiat pour notre champ d'application.

**Les textures** : elles permettent de caractériser les propriétés des matériaux présents dans l'image. Leurs calculs sont basés sur des analyses statistiques, structurelles ou spectrales [Liu *et al.*, 2009]. Cependant comme dans le cas de l'analyse des formes, l'utilisation de textures à des fins de caractérisation semble délicate dans le domaine de l'animation. En effet, les textures utilisées sont fortement liées aux techniques d'animation. Ces techniques et la diversité des textures utilisées pour construire les éléments du film traduisent la volonté artistique de l'auteur. On voit sur la figure 3.2 la grande variabilité des textures utilisées pour


2D or not 2D


La bouche cousue


Louise


Atama Yama


Casa


Le roman de mon âme


Hello


Moznosti dialogu


King Kong hr3

FIGURE 3.1 – Différentes formes de visages de quelques personnages d'animation.

composer les visages des personnages. Les matériaux utilisés caractéristiques de ces textures sont très nombreux et vont de la synthèse numérique à la pâte à modeler en passant par l'utilisation d'objets divers et variés comme des pâtes alimentaires, des légumes ou des ferrailles (écrou, vis, etc.). Dans ces conditions ce descripteur ne permettra probablement pas de reconnaître et nommer des objets à l'intérieur d'une image.

Dans le cas des séquences d'images la prise en compte de l'aspect temporel des images permet de caractériser les différents mouvements présents dans le film. **L'analyse globale** et **l'analyse locale** du mouvement sont des approches intéressantes pour l'analyse des films en général.

**Le mouvement global** : L'analyse du *mouvement global*, on l'a vu précédemment, consiste généralement à caractériser les mouvements de caméra [Bouthemy *et al.*, 1999] ou [Duan *et al.*, 2004]. Cette information de mouvement permet de localiser les passages importants de la séquence comme par exemple le fait de focaliser l'attention des spectateurs (arrêt sur une scène précise, puis zoom sur le visage d'un personnage). Ces techniques cinématographiques (zoom, travelling, etc.) sont généralement utilisées dans l'animation d'objets mais se retrouvent aussi quelques fois dans des dessins animés. Cette caractéristique ne nous a pas semblé très discriminante pour la description du contenu d'un film, elle ne sera donc pas

(a) Rusalka (b) Nos Adieux... (c) Amerlock

(d) Bunny (e) Histoire... (f) Ex-enfant

(g) Crac (h) How to cope... (i) Ferrailles

FIGURE 3.2 – Différentes textures composant les visages de personnages d'animation. Technique d'animation correspondant aux images :*(a) : peinture sur verre, (b) Nos Adieux au Music Hall : animation de pâtes alimentaires, (c) : pâte à modeler, (d) : ordinateur 3D, (e) Histoire Extraordinaire De Mme Keeskemet : dessin sur cellulos, (f) : écran d'épingles, (g) : dessin au crayon pastel, (h) How to cope with death : ordinateur 2D, (i) : animation d'objets.*

envisagée dans ce manuscrit.

**Le mouvement local** : La deuxième direction d'analyse est la caractérisation du *mouvement local ou mouvement des objets* [Laptev, 2005, Trucco et Plakas, 2006]. Cette information est intéressante pour faire du suivi d'objet mais là encore si le but est de caractériser sémantiquement ces mouvements (courir, marcher, voler, sauter, etc.) alors il devient difficile d'obtenir cette reconnaissance à partir des caractéristiques du mouvement seul comme cela peut être fait pour reconnaître des séquences où le mouvement local est déterminant, comme par exemple certaines séquences d'athlétisme [Ramasso, 2007]. En effet, la faible connaissance que l'on peut attacher à ces mouvements tant ils présentent de variabilité dans le domaine de l'animation ne permet pas leur caractérisation. Par exemple dans la célèbre série de cartoons américains produits par le studio Warner Bros (Bip Bip et Coyote, Speedy Gonzales, etc.) les personnages ont généralement des mouvements non compatibles avec les lois physiques du monde réel. Par conséquent, cette approche ne sera pas envisagée dans le cadre de notre étude. Nous y reviendrons à la fin de ce manuscrit dans les perspectives de nos travaux.

Parmi l'ensemble des possibilités technologiques offertes (description de scène, détection

d'objets, de personnes et de visages, etc.) un grand nombre de méthodes robustes fonctionnent bien en raison des hypothèses faites sur les données traitées. Par exemple, dans les vidéos sportives comme un match de football, les indicateurs de couleur, de forme et de texture sont très intéressants à exploiter et permettent de caractériser assez finement les éléments composant les images (terrain, joueurs) compte tenu du peu de variabilité de ces descripteurs associés aux concepts recherchés. De plus, les hypothèses faites sur les déplacements et les mouvements des objets permettent par exemple de retrouver l'action d'un joueur ou un tir au but. Malheureusement, un grand nombre de ces hypothèses sont remises en question dans le contexte des films d'animation. De ce fait, l'application des approches retrouvées dans la littérature du domaine est vite limitée car confrontée au passage des descriptions bas niveaux à des caractéristiques plus complexes.

Dans nos travaux, nous nous intéressons donc plus particulièrement à des caractérisations globales des films ; ainsi, nous cherchons à obtenir des éléments comme l'atmosphère, à partir d'informations utilisant les aspects couleurs véritablement spécifiques aux films d'animation et à partir d'information comme l'activité issue des séquences d'images.

### 3.1.2 L'existant pour les séquences d'animation

Nous présentons dans cette section un certain nombre de travaux développés au LISTIC sur la caractérisation des séquences d'animation, point de départ des travaux de thèse présentés dans ce mémoire. Dans l'objectif de caractériser les films d'animation à partir d'une information uni-modale, les travaux de Bodgan Ionescu [Ionescu, 2007] ont permis d'obtenir un ensemble de descripteurs symboliques basés sur la couleur et sur l'activité dans les séquences d'animation.

#### 3.1.2.1 Les descripteurs basés sur la couleur

Dans [Ionescu, 2007], une description statistique des couleurs prédominantes est calculée à partir d'un résumé de la séquence vidéo. Dans ces travaux, la distribution des couleurs est caractérisée par un *histogramme global pondéré des couleurs*. Cet histogramme sert ensuite de point de départ pour le calcul d'un certain nombre de descripteurs symboliques des couleurs dans l'analyse sémantique des séquences d'images.

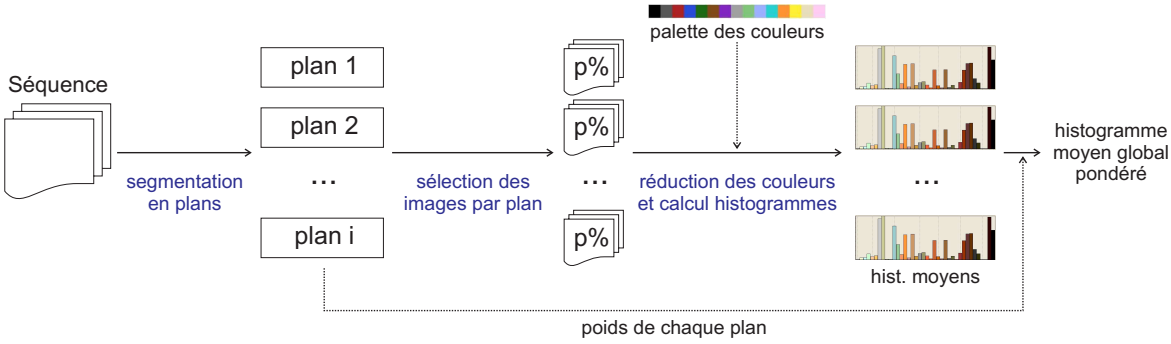



FIGURE 3.3 – Le calcul de l'histogramme global pondéré.

La méthode est illustrée par la figure 3.3 dont voici le détail des étapes :

- **le découpage en plans** : dans un premier temps la séquence est segmentée en plans vidéo par la détection des transitions vidéo du type : "cuts", "fades", "dissolves" et les changements brefs de couleurs SCC qui sont des effets de couleurs particuliers, spécifiques aux films d'animation (voir figure 2.1.c)). Les plans vidéo sont déterminés par la détection des transitions vidéo obtenues selon la méthode présentée dans [Ionescu *et al.*, 2005a]. Cette étape permet d'enlever les informations peu pertinentes du point de vue de la couleur comme par exemple les images de transition, les images noires, les plans trop courts qui sont peu visibles, etc.

- **le calcul du résumé** : un résumé de la séquence est calculé de manière automatique pour réduire la redondance temporelle. Dans ce résumé, chaque plan vidéo de la séquence est représenté par un pourcentage $p\%$ de ses images, centré sur le milieu du plan. En effet, il y a une très forte probabilité pour que l'action importante d'un plan se déroule en son milieu. Le meilleur compromis entre le temps de calcul et la qualité de la distribution globale des couleurs est obtenu empiriquement en prenant $p\% \in [15\%, 20\%]$.

- **la réduction des couleurs** : elle est appliquée sur une image sous-échantillonnée compte tenu de la quasi-invariance de l'histogramme couleur d'une image à un sous-échantillonnage spatial, c'est-à-dire en divisant la taille de l'image par un facteur k donné (k=4). Cette première étape permet de diminuer la complexité des calculs et donc de diminuer les temps d'exécution. De plus, la couleur des images numériques est habituellement représentée en utilisant 24 bits, soit plus de 16 millions de couleurs possibles. Ce nombre de couleurs (ou *"bin"* ou batons de l'histogramme) est bien trop élevé et une réduction préalable des couleurs est indispensable pour traiter efficacement les images. La méthode utilisée dans [Ionescu, 2007] consiste à associer chaque couleur de l'image à une couleur de référence contenue dans une palette de couleur fixe. La palette Webmaster [Visibone, 2009] de 216 couleurs est utilisée ici car elle présente l'avantage de décrire de manière textuelle chaque couleur, en terme de Teinte, de Clarté, etc. L'algorithme d'association des couleurs de l'image à celles de la palette est basé sur la diffusion d'erreurs et permet de conserver une bonne qualité visuelle de l'image après réduction des couleurs [Ionescu *et al.*, 2005b].

- **les histogrammes moyens** : ils sont calculés à partir des couleurs de chacune des images composant le résumé. Cet histogramme est une mesure de la distribution globale des couleurs du plan et a pour expression :

$$\bar{h}_i(c) = \frac{1}{N^i_{img}} \cdot \sum_{j=1}^{N^i_{img}} h_{i,j}(c) \tag{3.1}$$

où $N^i_{img}$ est le nombre d'images retenues dans le plan $i$ soit $p\%$ de ses images, $h_{i,j}(c)$ est l'histogramme couleur de l'image $j$ du plan $i$ et $c$ est l'indice des couleurs dans la palette "Webmaster" [Visibone, 2009], $c = 1, ..., 216$. Les histogrammes $h_{i,j}(c)$ sont calculés pour les images sous échantillonnées spatialement et avec les couleurs réduites. Les valeurs ainsi obtenues pour les histogrammes moyens sont normalisées entre 0 et 1 et représentent le pourcentage d'apparition des couleurs à l'intérieur de chaque plan.

– **l'histogramme global pondéré** de la séquence est la somme pondérée de tous les histogrammes moyens de chaque plan vidéo :

$$h_{seq}(c) = \sum_{i=1}^{N_{plans}} \bar{h}_i(c) \cdot \omega_i \tag{3.2}$$

où $\bar{h}_i(c)$ est l'histogramme moyen du plan $i$ avec $i = 1, ..., N_{plans}$ et $N_{plans}$ le nombre total de plans et où $c$ est l'indice des couleurs dans la palette "Webmaster" de 216 couleurs. La pondération ($\omega_i$) de chacun des histogrammes moyens dépend de la longueur du plan vidéo $i$ et vaut :

$$\omega_i = \frac{N^i_{img}}{N_{film}} \tag{3.3}$$

où $N^i_{img}$ est le nombre d'images du plan $i$ et $N_{film}$ est le nombre d'images de la séquence entière. Les valeurs de l'histogramme global pondéré, $h_{seq}(c)$, correspondent au pourcentage d'apparition de chaque couleur $c$ de la palette utilisée dans la séquence. Ce sont des valeurs positives qui sont calculées de telle sorte que leur somme soit égale à 1.

L'utilisation d'un histogramme couleur global pondéré permet de caractériser globalement la distribution des différentes couleurs dans la séquence. Cette caractérisation est importante dans les films d'animation et est motivée par le fait que les films d'animation utilisent généralement une palette de couleurs réduite propre à chaque film, sorte de signature couleur. De plus, une analyse de cet histogramme permet d'obtenir des caractéristiques d'un plus haut niveau sémantique.

Ces caractéristiques sont calculées à partir de l'histogramme couleur global pondéré dont voici quelques exemples :

Le *coefficient de couleurs claires*, $P_{claires}$, est la proportion des couleurs claires présentes dans la séquence. Il est facile de retrouver ces couleurs dans la palette "Webmaster" en utilisant la description textuelle associée. Ici il suffit que le nom de ces couleurs contienne des mots comme *"light"* ou *"pale"*. Ainsi $P_{claires}$ est défini par :

$$\begin{aligned} P_{claires} &= \sum_{c=1}^{216} h_{seq}(c)|_{\{Mot_{claire} \subset Nom(c)\}} \\ Mot_{claire} &\in \{\text{"}light\text{"}, \text{"}pale\text{"}, \text{"}white\text{"}\} \end{aligned} \tag{3.4}$$

où $c$ est l'indice d'une couleur et $Nom(c)$ est l'opérateur qui retourne le nom associé à la couleur d'indice $c$.

Le *coefficient de couleurs foncées*, $P_{foncées}$, représente le rapport des couleurs sombres présentes dans la séquence. Une couleur est considérée comme étant foncée si son nom contient l'un des mots suivants : "dark", "obscure", ou "black".

En raison du côté artistique des films d'animation et de l'utilisation réfléchie des couleurs par l'auteur, un certain nombre de règles issues du domaine de la peinture peuvent être utilisées pour caractériser des émotions ou sensations transmises par son œuvre [Itten, 1974]. Il est bien connu que certaines couleurs sont considérées comme dégageant une certaine chaleur, ou au contraire une sensation de froid. Sur ce principe et en correspondance avec la roue des couleurs de Itten, des descripteurs sont définis et permettent de quantifier ces sensations.

Le *coefficient de couleurs chaudes*, $P_{chaudes}$, est la proportion de couleurs chaudes présentes dans la séquence. Les couleurs considérées comme étant chaudes sont les couleurs appartenant à l'ensemble $\Gamma_{chaud} = \{$*"Yellow", "Orange", "Red", "Yellow Orange", "Red Orange", "Red Violet", "Magenta", "Pink" and "Spring"*$\}$ (voir figure 3.4). Une couleur de la séquence est considérée comme chaude si son nom contient l'un des mots de l'ensemble $\Gamma_{chaud}$.

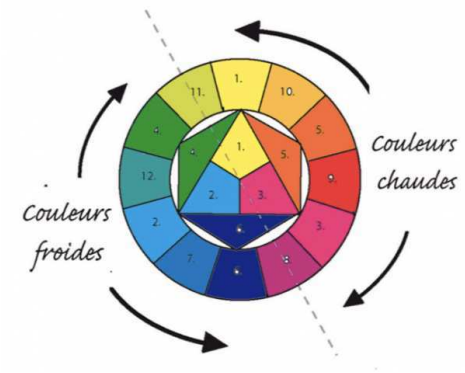



FIGURE 3.4 – Les couleurs chaudes et froides sur la route de Itten

D'autres descripteurs sont calculés à partir de la roue de Itten, comme la présence de couleurs adjacentes ou la présence de couleurs complémentaires. Pour voir les méthodes de calcul de ces descripteurs voir le chapitre 7 dans [Ionescu, 2007].

Toutefois, ces descriptions sont purement numériques et il est difficile d'apprécier cette valeur quantitative. Par exemple, lorsque l'on n'est pas spécialiste du domaine de l'animation, une valeur numérique de descripteur ne permet pas d'exprimer une sensation perçue. Une proportion de couleurs chaudes égale à 0.33 ne permet pas de se faire une idée qualitative de cette proportion. Est-ce que cette proportion est importante ou pas pour de tels films ? Pour répondre à cette question, l'approche envisagée dans [Ionescu, 2007] est de transformer cette valeur numérique en une valeur symbolique par l'utilisation d'ensembles flous. En effet, la formalisation floue permet la conversion entre les mesures numériques et les expressions linguistiques proches de notre mode de perception. Ainsi le concept de proportion de couleurs chaudes est décrit en utilisant trois variables linguistiques illustrées par les symboles suivants : *"présence Faible de couleurs chaudes"*, *"présence Moyenne de couleurs chaudes"* et *"présence Haute de couleurs chaudes"*. La signification floue de chaque symbole est traduite par sa fonction d'appartenance : $\mu_{Faible}$, $\mu_{Moyenne}$ et $\mu_{Haute}$ et est illustrée par la figure E.6. Le principe est le même pour le calcul des valeurs symboliques des autres descripteurs.

De plus, le formalisme flou permet la combinaison de ces descripteurs par l'utilisation de règles de combinaison et permet de construire de nouvelles expressions linguistiques (inférences floues) comme par exemple le concept de *froideur*. L'approche floue et ses concepts seront présentés dans le chapitre 5 consacré à la fusion d'information.

Finalement, ces descripteurs couleurs seront utilisés dans le chapitre 5 consacré à la fusion d'information pour caractériser les films à partir de la modalité image.

#### 3.1.2.2 Les descripteurs basés sur l'analyse des plans vidéo

De façon similaire à ce qui a été fait pour la caractérisation des couleurs dans les séquences d'animation, les travaux dans [Ionescu, 2007] fournissent des descripteurs obtenus à partir de la distribution des plans vidéo.

Pour caractériser la distribution des plans vidéo, un indicateur nommé $\zeta_T(i)$ est calculé à partir de la détection des transitions et représente le nombre de changements de plan survenus dans une plage de durée $T$ à partir de l'image à l'instant $i$. Cet indicateur est lié à la structure temporelle de la séquence et permet de calculer deux descripteurs : le *rythme de la séquence* et la *mesure de l'action*.

**Le rythme de la séquence** ($\bar{v}_T$) représente le nombre moyen de changements de plan dans une fenêtre de temps $T$ (valeur fixée empiriquement à $T = 5s$). Le paramètre $\bar{v}_T$ est lié au déroulement temporel de la séquence. Plus il y a de l'activité, c'est-à-dire de changements de plan par unité de temps $T$ plus $\bar{v}_T$ est élevé. Cela traduit donc un rythme de la séquence d'animation qui peut être décrit par trois valeurs linguistiques : *"rythme lent"*, *"rythme moyen"* et *"rythme rapide"*. La signification floue de chaque symbole est illustrée par sa fonction d'appartenance floue. La partition floue de l'univers de discours, $\bar{v}_{T=5s}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\mu_{lent}$, $\mu_{moyen}$ et $\mu_{rapide}$ (voir la figure 3.5).

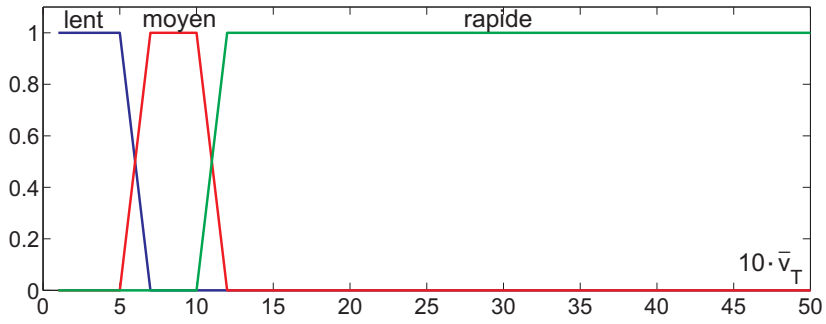



FIGURE 3.5 – La partition floue de l'univers de discours $\bar{v}_T$ déterminée par les fonctions d'appartenance floues : $\mu_{lent}$ (bleu), $\mu_{moyen}$ (rouge) et $\mu_{rapide}$ (vert) (l'axe des ordonnées correspond au degré d'appartenance).

**L'action moyenne de la séquence** ($R_{action}$) représente la proportion de passages où l'activité est significative par rapport au reste de la séquence. L'hypothèse faite dans

[Ionescu, 2007] est que les passages du film où il y a beaucoup de changements de plan sont des passages où l'action est significative. Cette hypothèse est justifiée par le fait que la relation entre la fréquence des changements de plan et l'action est très souvent utilisée dans les techniques de génération automatique de résumés de séquences, comme les "bande-annonces". A partir de cette hypothèse, *B.Ionescu* construit un signal binaire fonction du temps, défini par :

$$f_{action}(i) = \begin{cases} 1 & \text{si } \zeta_T(i) > \bar{v}_T \\ 0 & \text{sinon} \end{cases} \tag{3.5}$$

qui représente les passages où l'action est significative par rapport à l'ensemble de la séquence. C'est-à-dire que l'action est considérée comme significative si le nombre de changements de plan par unité de temps $T$ est supérieur à la moyenne $\bar{v}_T$ de $\zeta_T(i)$ calculée sur l'ensemble de la séquence. Enfin, quelques post-traitements sont appliqués sur la fonction $f_{action}(i)$ pour éliminer les segments trop courts et pour fusionner les segments très proches, obtenant ainsi la fonction binaire action $F_{action}(i)$.

Une description globale de l'action est ensuite calculée à partir de la fonction action $F_{action}(i)$. Le paramètre $R_{action}$ représente le pourcentage de segments d'action par rapport à la séquence entière :

$$R_{action} = \frac{T_{action}}{T_{film}} \tag{3.6}$$

où $T_{action}$ est la durée totale des segments d'action et $T_{film}$ est la durée totale du film. Comme pour les autres descripteurs, le concept d'action est décrit par trois valeurs linguistiques, *"action faible"*, *"action moyenne"* et *"action élevée"* est calculé à partir de cette mesure.

#### 3.1.2.3 Limites de ces descripteurs

Nous venons de voir à partir des travaux de thèse de Bogdan Ionescu qu'un certain nombre de descripteurs bas niveau sont extraits des séquences d'images. Ces descripteurs sont de natures bien différentes et correspondent d'une part à une description globale des couleurs et de leurs rapports "artistiques" dans les images et d'autre part à une description de l'activité et du rythme dans la séquence vidéo. Ces descripteurs sont calculés à partir de l'analyse des plans vidéo qui correspond au premier niveau d'analyse de la pyramide des niveaux sémantiques des documents vidéo (voir figure 1.2). Cette analyse va donc bien dans le sens proposé par [Davenport *et al.*, 1991] et donne de bons résultats sur les films construits sur ce modèle. Cependant, il apparaît trois problèmes majeurs dans l'obtention de ces descripteurs lorsque l'on veut traiter l'ensemble de la base d'animation de CITIA :

- L'hypothèse de construction du document vidéo à partir de plans vidéo est habituellement vérifiée lorsque l'on traite des films naturels, longs métrages (au moins 90 minutes) ou courts métrages (moins de 45 minutes). Mais cette hypothèse est inappropriée sur de nombreux films d'animation dont nous disposons. En effet, 44% des films de la base sont considérés comme de très courts métrages (moins de 5 minutes, voir figure B.2). Ainsi, la construction du film d'animation est vite limitée à quelques plans vidéo comme c'est généralement le cas dans les films en pâte à modeler où les personnages et les objets apparaissent et disparaissent de la scène continuellement. Les caractéristiques de couleurs extraites ne sont plus du tout pertinentes si il y a peu de plans vidéo (on

rappelle en effet que les descripteurs couleurs reposent sur la construction d'un résumé statique obtenu à partir des plans voir la figure 3.3).

– L'activité dans la séquence d'animation est mesurée à partir de la distribution des transitions. Or la sensation de rythme dans une séquence ne se traduit pas seulement par l'utilisation de plans différents mais passe aussi par la sensation de l'activité à l'intérieur de ces plans. Par conséquent cette manière de procéder ne prend pas en compte l'activité intra-plan qui est liée aux changements entre les images d'un même plan. Il semble alors plus judicieux de mesurer cette activité directement à partir des changements de contenu des images composant le film d'animation.

– Enfin, une troisième limite concerne la mesure de l'action globale. En effet, le paramètre $R_{action}$ représente le pourcentage de segments d'action par rapport à la séquence entière. Or les segments d'action sont définis comme les passages où le nombre de changements de plan par unité de temps $T$ est supérieur à la moyenne $\bar{v}_T$ de $\zeta_T(i)$ calculée sur l'ensemble de la séquence. Par conséquent $R_{action}$ est une mesure relative au film et ne permet pas une comparaison absolue des films entre eux.

Finalement, pour s'affranchir de ces défauts, nous proposons un nouvelle méthodologie pour le calcul des descripteurs de couleur et de rythme. Cette approche basée sur la mesure des changements de contenu s'opérant dans les images évite la détection des transitions et propose des mesures d'activité absolue.

## 3.2 Propositions

Nous avons vu que l'approche par détection de plans a ses limites sur certains films d'animation et nous souhaitons nous affranchir de cette segmentation en plans.

### 3.2.1 Les objectifs

L'objectif principal est de conserver les descripteurs issus des travaux de la thèse de Bogdan Ionescu, tout en rendant les méthodes de calcul généralisables à l'ensemble des films de la base de CITIA (en particulier dans le cas des films monoplan). Nous proposons deux améliorations :

– Réduire la séquence vidéo à un ensemble d'images représentatives du film, dans lequel la redondance de contenu est fortement réduite. Ce condensat d'images servira de "résumé statique", point de départ pour le calcul des histogrammes couleurs (voir la figure 3.3).

– Mesurer l'activité et le rythme dans la séquence vidéo à partir des changements de contenu s'opérant à l'intérieur des images. En effet, nous partons de l'hypothèse que l'activité dans les films d'animation est liée à la fréquence du changement de contenu.

### 3.2.2 Notre approche

Les deux objectifs présentés ci dessus ont en commun d'être basés sur une analyse du contenu des images. Cette analyse consiste à retrouver parmi toutes les images de la séquence vidéo, les images dont le contenu est suffisamment différent du reste de la séquence. Cela impose de mesurer une dissimilarité entre les images. En effet, nous ne cherchons pas à savoir où se trouve les différences de contenu, ni à savoir quelle est la nature de ce contenu. Nous voulons retrouver les images qui d'un point de vue global (celui du spectateur) ont une différence significative entre elles. De plus, il semble assez évident qu'il y a beaucoup plus d'images différentes (c'est-à-dire dont le contenu est différent) dans les passages du film où l'action est élevée que dans les passages où il y a peu d'action. Par conséquent, la densité d'images différentes servira à mesurer l'activité inter et intra-plans de la séquence. Enfin, l'approche que nous proposons permet de s'affranchir de l'étape de détection des transitions, ce qui lui confère une plus grande généralité.

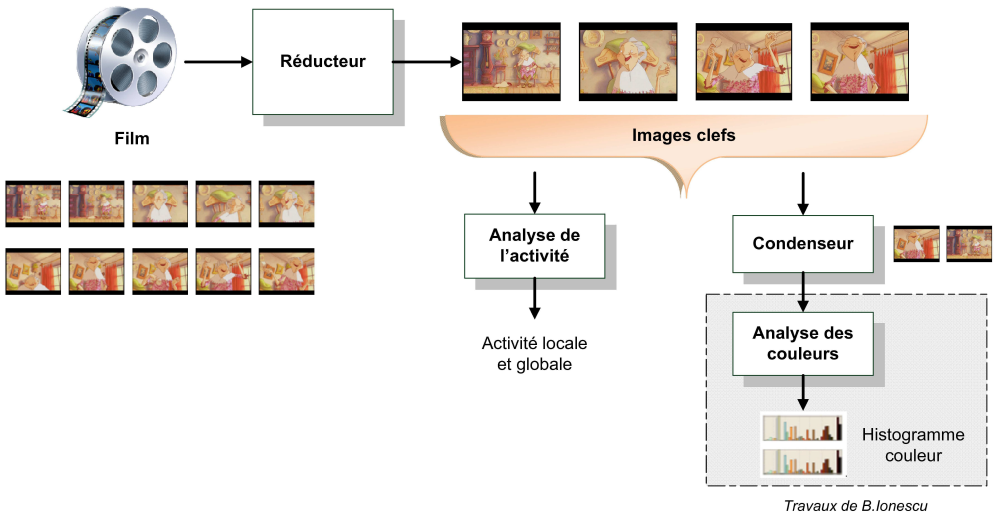



FIGURE 3.6 – Vision globale du système

L'organisation générale de notre approche est présentée sur la figure 3.6. La première étape nommée **Réducteur** permet de réduire le nombre d'images de la séquence vidéo en supprimant la redondance entre les images successives, tout en gardant l'essentiel de l'information contenue dans le film. Son rôle est double. D'une part elle permet d'extraire dans le temps, un ensemble d'***"images clefs"*** dont le contenu est globalement différent d'une image à l'autre. L'analyse de la distribution temporelle de ces images clefs permettra de mesurer l'activité dans la séquence vidéo (étape d'**analyse de l'activité**). D'autre part, c'est un pré-traitement pour l'étape du **condenseur** qui permet la création du condensat (sorte de résumé global sans aucune sémantique) en ne sélectionnant qu'un nombre limité d'images pour le calcul et l'**analyse des caractéristiques couleur**.

Dans les sections suivantes, nous détaillerons les différentes étapes présentées ci-dessus. Nous commencerons par détailler l'étage du réducteur puis nous verrons comment caractériser l'activité et le rythme dans les films d'animation pour finir par la méthode permettant de

générer le condensat préalable à l'extraction des caractéristiques couleur.

### 3.2.3 La détection du changement de contenu

Pour la détection des plans, deux images consécutives sont généralement comparées entre elles via une métrique qui traduit la ressemblance ou dissemblance de contenu entre ces images. Calculée entre 2 images successives, cette différence reste généralement faible, même en présence d'un mouvement de caméra d'amplitude moyenne, du déplacement des éléments de la scène ou d'un changement dans le fond de l'image. Une valeur importante de cette mesure correspond la plupart du temps à un changement de plan, ou à un effet spécial. Ainsi si cette comparaison dépasse un seuil alors un changement de plan *"shot break"* est déclaré. Replacé dans le contexte des films d'animation et à la lumière de ce qui a été précisé plus tôt, nous pouvons avec ce principe détecter :

- Un changement de plan.
- Une différence nette entre les deux images comparées qui peut être liée à un effet particulier. Par exemple un changement bref de couleur (SCC) (voir figure 2.1).

Or une séquence vidéo contient beaucoup d'images similaires consécutives. Une seconde de film est équivalente à 25 images et dans une scène continue où l'activité n'est pas très élevée il n'y a pas de gros changements durant cette seconde. Il y a donc au moins 25 images qui sont quasi semblables, d'où une forte redondance dans le contenu. Ainsi l'idée de comparer les images une à une consécutivement ne permet pas de faire apparaître des différences de contenu mais seulement des discontinuités nettes. Nous avons donc mis en place un algorithme capable de détecter les changements de contenu dans la vidéo en nous inspirant des travaux de **Tong** [Lu et Suganthan, 2004].

#### 3.2.3.1 L'algorithme à accumulation de différences

Dans l'objectif de détecter des transitions graduelles comme les fondus enchaînés "fade" et "dissolve", qui sont deux types de transition qui ne se traduisent pas par un changement brutal du contenu entre images successives (voir figure 1.3), **Tong** a introduit un algorithme à accumulation de différences dont la particularité est sa capacité à mémoriser les changements entre images consécutives. Nous avons donc adapté cet algorithme à notre problème et aux films d'animation.

**3.2.3.1.1 Principe** On suppose disposer d'une mesure de différence entre images traduisant la dissemblance de contenu entre ces images. Pour arriver à extraire d'une séquence vidéo un ensemble d'images significativement distinctes, une solution consiste à accumuler progressivement les différences entre images successives. Lorsque l'accumulation de ces différences dépasse un seuil, cela signifie que les images correspondant au début et à la fin de l'accumulation peuvent être considérées comme distinctes. La remise à zéro du système ainsi que l'itération de ce mécanisme permet d'extraire un ensemble d'images clefs.

La figure 3.7 illustre ce principe. La première image de la séquence est utilisée comme image de référence. Ensuite les images successives (en abscisse) sont comparées à cette image

de référence, les différences sont accumulées jusqu'à ce que leur nombre (en ordonnée) dépasse le seuil (en vert) fixé empiriquement. L'image pour laquelle la valeur des différences accumulées dépasse le seuil est marquée comme significativement différente de l'image prise comme référence (on parle aussi d'image clef). Elle devient à son tour l'image de référence pour la suite du traitement. Le processus s'arrête lorsque la dernière image de la séquence est analysée. Il faut noter que ce mécanisme permet à la fois de détecter les changements de plan et les évolutions progressives du contenu des images.

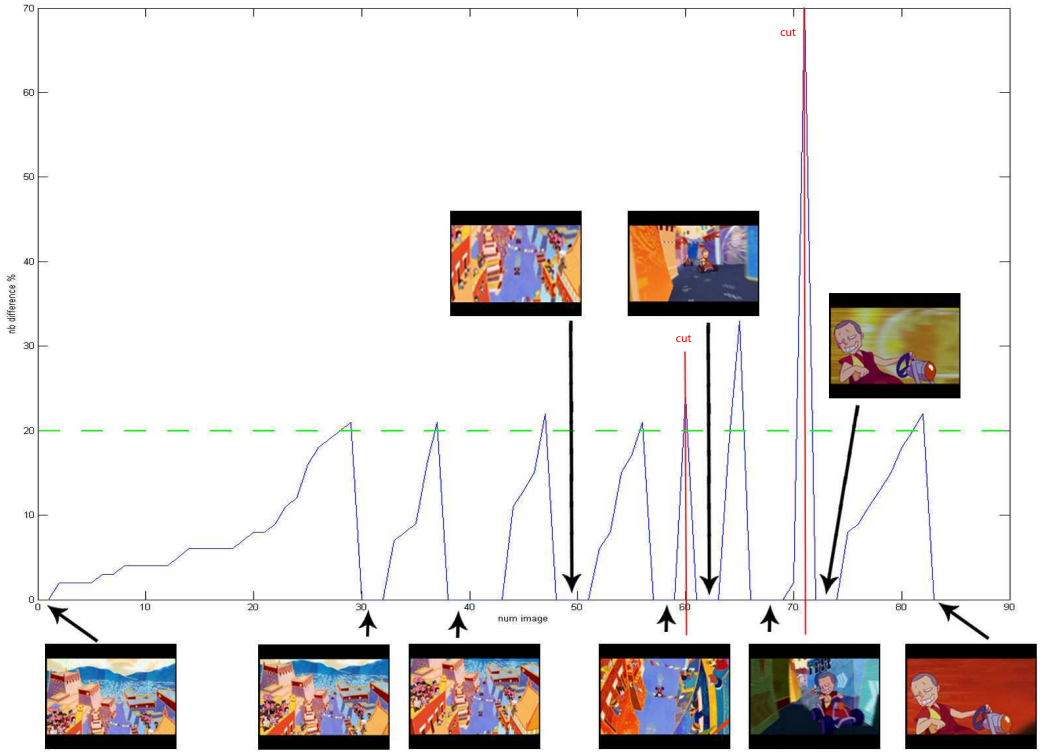



FIGURE 3.7 – Graphique illustrant le principe de l'accumulation de différences *(Le pourcentage de différences est en ordonnée et le numéro de l'image en abscisse. Le seuil est représenté en vert)*

**3.2.3.1.2 Fonctionnement** On l'a vu précédemment l'algorithme de Tong est basé sur une mesure de similarité entre images. Pour mettre en forme et comparer les images entre elles plus rapidement, l'auteur utilise une méthode de découpage en blocs. Chaque image est découpée en un ensemble de N*N blocs de pixels de tailles égales dont les valeurs moyennes (couleur des pixels composant le bloc) permettent de créer une matrice réduite (de N*N éléments) de l'image. La comparaison de deux images se fait donc en comparant les matrices réduites des images (élément à élément) en utilisant un seuil $T_b$ fixé par l'utilisateur et une distance dans l'espace RGB. Ce seuil fixe la valeur de la distance entre deux éléments au delà de laquelle ces deux éléments sont considérés comme différents.

L'algorithme 1 présente le fonctionnement du système. La première image du film $Im$ (d'index $i = 1$) est extraite de la vidéo en utilisant la fonction **Film.getImageAt()** puis

**Algorithme 1:** Accumulation de différences

**input** : Un film ($Film$)
**output** : La liste des images clefs ($ListNumKeyFrame$)

```
1  begin
2      ListNumKeyFrame ← 0;
3      Mat_state ← 0;
4      Im ← Film.getImageAt(1);
5      I_ref ← ReductInBlocs(Im);
6      for i ← 2 to Film.nbImage() do
7          Im ← Film.getImageAt(i);
8          I_cur ← ReductInBlocs(Im);
9          //compare I_ref et I_cur puis marque les différences dans Mat_state
10         foreach element(m,n) in matrix Mat_state(m,n) do
11             //si pas déjà marqué différent
12             if Mat_state(m,n) ≠ State_different then
13                 Mat_state(m,n) ← CompareAndGetState(m,n,Mat_state,I_ref,I_cur);
14             end
15         end
16         //si le nombre de différences excède le seuil Td
17         if Td ⩽ NumberOfDifferentBloc(Mat_state) then
18             //ajout d'une image clef
19             ListNumKeyFrame.add(i);
20             Mat_state ← 0;
21             I_ref ← I_cur ;
22         end
23     end
24 end
```

elle est réduite en blocs, comme expliqué ci-dessus, via la fonction **ReductInBlocs()**. Cette image réduite est ensuite copiée dans la matrice de référence $I_{ref}$ et devient l'image de référence (ligne 5). L'image suivante d'indice $i$ est à son tour réduite puis stockée dans la matrice courante $I_{cur}$ (ligne 8).
Les deux matrices précédentes $I_{cur}$ et $I_{ref}$ sont comparées (lignes 10 à 15) via la fonction **CompareAndGetState()** en comparant leurs éléments deux à deux. Le résultat de ces N*N comparaisons est retranscrit dans la matrice d'état $Mat_{state}(m,n)$ aussi appelée accumulateur. Chaque élément de cette matrice correspond à l'état de la comparaison entre $I_{cur}(m,n)$ et $I_{ref}(m,n)$ (ligne 13). De plus, les éléments d'indice $(m,n)$ de la matrice $Mat_{state}$ sont mis à jour seulement lorsque l'état de la comparaison précédente (entre l'image de référence et l'image d'indice $i-1$) n'est pas l'état $State_{different}$ (ligne 12). Ainsi, un élément de la matrice d'état marqué comme différent le restera jusqu'à la remise à zéro de l'accumulateur ($Mat_{state} \leftarrow 0$ ligne 19). Finalement cette remise à zéro est exécutée lorsque l'accumulation c'est-à-dire lorsque le nombre d'éléments marqués comme différents dans l'accumulateur ($State_{different}$) dépasse un seuil $Td$ (ligne 17). Le nombre d'éléments marqués comme différents ($Mat_{state}(m,n) == State_{different}$) est obtenu grâce à la fonction **NumberOfDif-**

**ferentBloc()** (ligne 17). Le seuil $Td$ représente le pourcentage d'éléments marqués comme différents dans $Mat_{state}$ et il est fixé empiriquement à 20%. Ainsi, lorsque le nombre de différences dépasse ce seuil, alors l'indice de l'image courante $i$ est enregistré dans la liste des images clefs **ListNumKeyFrame.add(**$i$**)** (ligne 18), la matrice d'état est remise à zéro (ligne 19) et l'image courante est prise comme nouvelle référence ($I_{ref} \leftarrow I_{cur}$ ligne 20). L'itération de ce processus s'arrête lorsque la dernière image de la séquence est analysée ($Film.nbImage$ ligne 6). Le principe de l'accumulation réside dans le fait que les éléments de la matrice d'état marqués comme différents le restent jusqu'à ce qu'une nouvelle image soit prise comme référence.

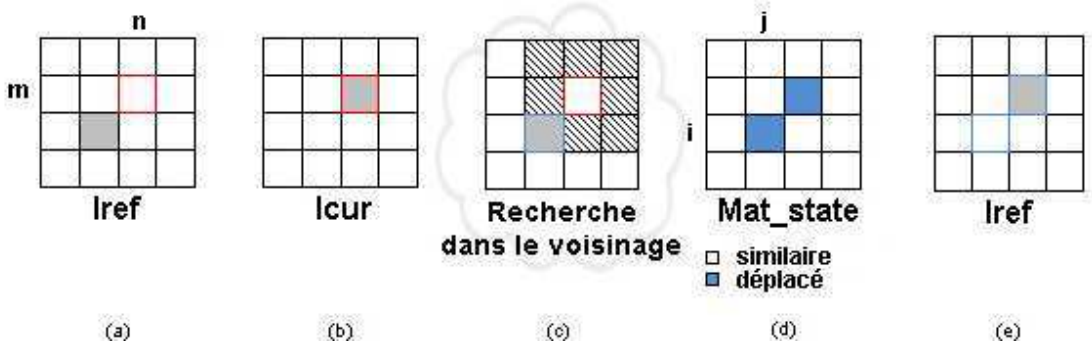



FIGURE 3.8 – Compensation du mouvement (recherche des blocs se déplaçant), (a) accumulateur, (b) image à analyser, (d) matrice d'état

Comme le mouvement des objets et de la caméra est important dans les films d'animation et spécialement dans les films monoplan, **la compensation du mouvement** utilisée par Tong est conservée dans l'algorithme. Imaginons que d'une image à l'autre un bloc de pixels se déplace d'une position (par exemple le bloc gris de l'image a. à l'image b. sur la figure 3.8). La comparaison entre les matrices $I_{ref}$ et $I_{cur}$ conduira à considérer les deux blocs (entourés de rouge en a. et b.) $I_{ref}(m,n)$ et $I_{cur}(m,n)$ comme différents alors qu'en réalité le contenu des images reste globalement le même, il s'est juste déplacé. Afin de palier ce problème l'algorithme va vérifier si l'élément $I_{cur}(m,n)$ ne s'est pas déplacé. Si le bloc analysé ($I_{cur}(m,n)$) se retrouve dans l'un de ses 8 plus proches voisins dans la matrice de référence $I_{ref}(m,n)$ (c'est le cas sur l'exemple c. de la figure 3.8 le bloc $I_{cur}(m,n)$ est retrouvé en $I_{ref}(i,j)$) alors les éléments en question ne seront pas marqués comme différents mais marqués comme déplacés dans la matrice d'état $Mat_{state}$ (bloc bleu sur la figure 3.8 d. et lignes 6 et 7 de l'algorithme 2). La matrice de référence $I_{ref}$ est mise à jour en recopiant de l'image courante $I_{cur}$ les deux éléments de départ et d'arrivée (sur la figure 3.8 e. $I_{ref}(m,n) \leftarrow I_{cur}(m,n)$ et $I_{ref}(i,j) \leftarrow I_{cur}(i,j)$ lignes 9 et 10 de l'algorithme 2). Finalement la mise à jour de l'accumulateur permet de suivre et de compenser ce déplacement.

Cependant il apparaît un deuxième problème avec cette solution. Imaginons que d'une image à l'autre l'objet n'a pas le temps de se déplacer d'un bloc à un autre mais qu'il ne s'est déplacé que d'une partie de bloc (voir figure 3.9). Ce mouvement partiel ne sera pas détecté par la compensation de mouvement et conduira à considérer l'élément comme différent. Pour remédier à ce problème la solution consiste à ne pas statuer immédiatement sur l'état du bloc mais à attendre pour vérifier si il y a un déplacement sur les images suivantes (sur les images $i+1$, $i+2$, ...). Ainsi durant cette période transitoire les blocs sont marqués

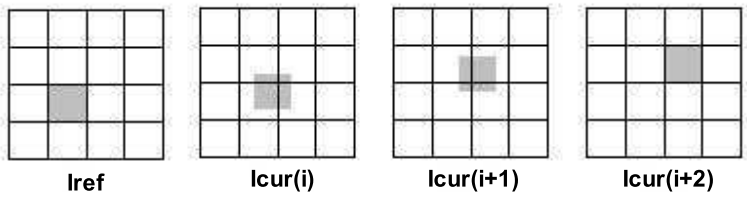



FIGURE 3.9 – Mouvement partiel d'un bloc au court du temps

dans l'accumulateur ($Mat_{state}$) comme temporairement différents. On considère qu'au bout de 4 images consécutives ($i+4$) le bloc doit avoir fini son mouvement partiel et donc être retrouvé parmi ses voisins. Sinon, il passe d'un état temporairement différent à un état différent.

Ce mécanisme d'attente se retrouve dans l'implémentation de la fonction *CompareAndGetState()* entre les lignes 20 et 26 de l'algorithme 2. Lorsque les éléments $I_{cur}(m,n)$ et $I_{ref}(m,n)$ sont différents (ligne 2) et que l'élément $I_{cur}(m,n)$ n'a pas été retrouvé dans le voisinage de $I_{ref}(m,n)$ (ligne 13) alors l'algorithme décrémente l'élément $Mat_{state}(m,n)$ de la matrice d'état (ligne 22). Chaque état de la matrice $Mat_{state}$ est en réalité un chiffre dont la signification est la suivante :

| État | $State_{different}$ | *temp3* | *temp2* | *temp1* | $State_{same}$ | $State_{move}$ |
|---|---|---|---|---|---|---|
| Numéro | -4 | -3 | -2 | -1 | 0 | 1 |

On peut noter que la différence d'index minimum séparant deux images clefs consécutive vaut 4. En effet, si l'image de référence a l'index $i$ et si les images courantes $i+1$, $i+2$, $i+3$, $i+4$ sont toutes différentes alors tous les blocs de la matrice d'état passeront par les états $temp1$, puis $temp2$, puis $temp3$ et enfin $State_{different}$. Finalement la première image clef aura l'index $i$ et la deuxième image clef aura l'index $i+4$. Notons également que ce mécanisme permet d'éliminer le bruit ou des effets spéciaux de courte durée comme les changements brefs de couleur.

#### 3.2.3.2 Améliorations et nouveautés

On l'a vu précédemment l'algorithme de Tong est basé sur une mesure de similarité entre images. Cette similarité est évaluée en comparant les matrices réduites des images (élément à élément) en utilisant un seuil et une distance dans l'espace RGB. Nous avons modifié cette étape car ses performances étaient très dépendantes du choix du seuil. En effet, il est difficile d'envisager un seuil fixe pour l'ensemble des films d'animation. De plus pour un même film, ce seuil est difficilement appréciable. Aussi, le principe de découpage en blocs a été conservé mais on extrait pour chaque cellule la valeur médiane vectorielle des pixels composant le bloc (voir figure 3.10.c), ceci afin de ne pas faire apparaître des fausses couleurs (voir figure 3.10.b) tout en atténuant le bruit dans l'image. En effet, certaines vidéos sont issues d'une numérisation de support comme les VHS et comportent beaucoup de bruit.

La comparaison des blocs a également été repensée. En effet, dans les travaux de Tong

**Algorithme 2:** CompareAndGetState($m$,$n$,$Mat_{state}$,&$I_{ref}$,$I_{cur}$)

```
begin
    if I_ref(m,n) ≠ I_cur(m,n) then
        //Chercher parmi les 8 voisins de I_ref(m,n) d'indices i,j l'élément I_cur(m,n)
        foreach i ∈ [−1;0;1] and j ∈ [−1;0;1] do
            if i ≠ 0 and j ≠ 0 then
                if I_ref(m+i,n+j) = I_cur(m,n) then
                    i ← m+i;
                    j ← n+j;
                    break;
                end
            end
        end
        if I_cur(m,n) a migré de la position (i,j) dans I_ref then
            //Marquer comme déplacé
            Mat_state(m,n) ← State_move;
            Mat_state(i,j) ← State_move;
            //Met à jour la matrice de référence
            I_ref(m,n) ← I_cur(m,n) ;
            I_ref(i,j) ← I_cur(i,j) ;
        else
            //temporairement différent
            Mat_state(m,n) ← Mat_state(m,n)-1;
            if Mat_state(m,n) ⩽ -4 then
                Mat_state(m,n) ← State_different;
            end
        end
    else
        Mat_state(m,n) ← State_same;
    end
    return Mat_state(m,n)
end
```

cette comparaison est effectuée à l'aide de la distance Euclidienne dans l'espace RGB. Or, les distances calculées dans cet espace ne reflètent pas la perception de proximité colorimétrique de l'homme. De plus, il est difficile dans cet espace de choisir un seuil fixe permettant de déterminer si deux blocs sont similaires ou non. Afin de supprimer cette contrainte du seuil et en raison des caractéristiques couleurs des films que nous traitons, nous préférons baser cette étape sur la comparaison des noms de couleur issus de la technique de *"color naming"*. Cette technique consiste à nommer les couleurs, c'est-à-dire que la couleur d'un bloc n'est plus représentée par un triplet de valeurs numériques représentant la couleur médiane mais par le nom de cette couleur médiane. Ainsi chaque image est représentée par une matrice de taille N*N dont chaque élément est un nom de couleur associé à la valeur colorimétrique médiane des pixels composant le bloc équivalent. La comparaison de deux images passe par la

(a) (b) (c)

FIGURE 3.10 – Illustration de la réduction d'un bloc de pixel en une valeur RGB. *(a) : Bloc de pixels à réduire, (b) : Réduction avec la moyenne fait apparaître une nouvelle couleur marron, (c) réduction avec la médiane vectorielle fait apparaître la couleur réelle majoritaire*

comparaison des noms de couleur des éléments (de même indice) de leurs matrices réduites. Le mécanisme d'association d'un nom à une couleur *"color naming"* est détaillé ci-dessous.

#### 3.2.3.3 La comparaison par *"color naming"*

On l'a vu précédemment chaque image est réduite (dans une matrice N*N) afin de simplifier les comparaisons. Mais la principale difficulté est d'établir la similarité ou la différence entre deux blocs. Une méthode généralement utilisée est de prendre une métrique dans un espace colorimétrique adéquate (La*b* par exemple). Le problème avec cette solution est qu'il faut choisir l'espace colorimétrique mais surtout qu'il faut introduire un seuil permettant de dire pour une distance calculée si les blocs sont similaires ou non. Afin de simplifier la tâche de comparaison et de supprimer ce seuil, l'idée proposée est de comparer les blocs comme le fait un être humain. C'est-à-dire en utilisant un ensemble de symboles de référence (en l'occurrence le nom des couleurs) qui permettent de pouvoir comparer plus facilement 2 couleurs entre elles. Avec cette méthode la comparaison entre 2 blocs devient très simple puisqu'elle est binaire (même nom de couleur ou noms différents). Afin d'exploiter cette propriété nous avons besoin d'un système capable pour une couleur donnée, de fournir son nom.

Il existe plusieurs dictionnaires de noms de couleur. Celui que nous avons retenu est le dictionnaire ISCC-NBS [Kelly et Judd, 1955] commanditée par le Inter-Society Color Cuncil (ISCC) à l'intention du National Bureau of Standards (NBS). Pour simplifier la description d'une couleur, l'ISCC-NBS a standardisé le nom de 267 couleurs du nuancier de Munsell. Chaque nom a donc une correspondance exacte avec une couleur bien déterminée dans ce système. Pour ce faire, les 10 termes de base suivants ont été retenus : "pink", "red", "orange", "brown", "yellow", "olive", "green", "blue", "violet", "purple". Puis 28 noms ont été créés à partir de ces termes, par combinaison en paire : "reddish orange" (orange tirant sur le rouge), "bluish green" (vert tirant sur le bleu), etc..., auxquels il faut ajouter les 3 noms "white", "gray" et "black" (respectivement blanc, gris et noir). Enfin des adjectifs ("very", "strong", "vivid",...) sont choisis pour traduire les nuances d'une teinte donnée (voir figure 3.11). Des travaux comme ceux de Menegaz [Menegaz *et al.*, 2007] utilisent d'autres dictionnaires comme par exemple celui proposé par Berlin et Kay [Berlin et Kay, 1969] qui se limite à 11 couleurs de base. Bien qu'un film d'animation utilise un jeu de couleurs (palette) réduit par rapport à la diversité des couleurs que l'on trouve dans les séquences naturelles, se limiter à 11 couleurs est beaucoup trop restrictif pour permettre une comparaison acceptable entre les images.

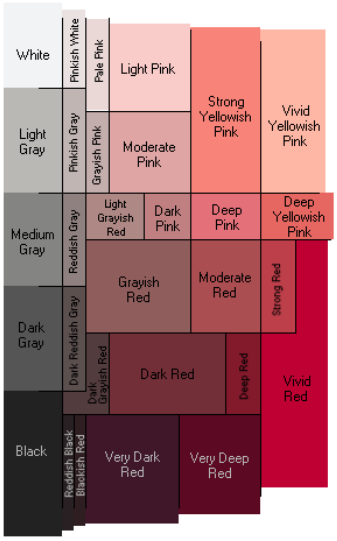



FIGURE 3.11 – Extraction de la table des couleurs ISCC-NBS de "Pink" et "Red"

Finalement utiliser une réduction à 267 termes est un bon compromis [Ionescu *et al.*, 2005b] car cela permet de rester fidèle aux couleurs d'origine de l'image tout en réduisant le nombre de comparaisons possibles. De plus, la notation ISCC-NBS tient à n'utiliser qu'un nombre restreint de termes directement explicites et à les combiner entre eux ce qui peut être utile pour une comparaison plus « intelligente » des couleurs. Notons également que la classification faite par l'ISCC-NBS fait que deux couleurs avec deux noms différents sont visuellement bien différentes. Finalement cette comparaison des couleurs est plus évoluée qu'une simple quantification de l'espace couleur car plus proche d'une approche humaine.

Le système de "color naming" utilisé est basé sur l'algorithme de [Mojsilovic, 2005] qui pour une couleur donnée recherche la couleur du dictionnaire qui visuellement semble la plus proche dans l'espace La*b*. Pour illustrer ce principe la couleur dont on cherche le nom (représentée par une étoile dans l'espace 3D CIELAB sur la figure 3.12) est comparée aux couleurs de référence du dictionnaire couleur en utilisant une distance basée sur la distance DeltaE76 (voir equation 3.14). Ainsi, l'algorithme calcule les distances (représentées par des flèches noires sur la figure 3.13) entre la couleur dont on cherche le nom (représentée par une étoile de couleur bleue sur la figure 3.13) et les couleurs de référence (représentées par des pavés colorés). La couleur dont on cherche le nom prend le nom de la couleur de référence dont elle est la plus proche c'est-à-dire dont la distance est la plus faible (dans l'exemple de la figure 3.13 la couleur analysée porte ainsi le nom "red blue"). Notons que la figure 3.13 n'est qu'une représentation 2D par projection le long de l'axe de luminosité (L) de l'espace CIELAB 3D. Les 267 distances sont bien sûr calculées dans l'espace CIELAB 3D. La figure 3.13 permet d'illustrer simplement le mécanisme de "color naming".

**3.2.3.3.1 Comparaison des couleurs basée sur leur nom** A partir de la table initiale du ISCC-NBS composée de 267 couleurs et du mécanisme de "color naming" nous sommes capables de comparer simplement deux couleurs entres elles. Si elles n'ont pas le même nom alors elles sont considérées comme différentes. Cependant, cette façon de procéder semble assez brutale. En effet, parmi les couleurs du dictionnaire beaucoup sont des variantes d'une

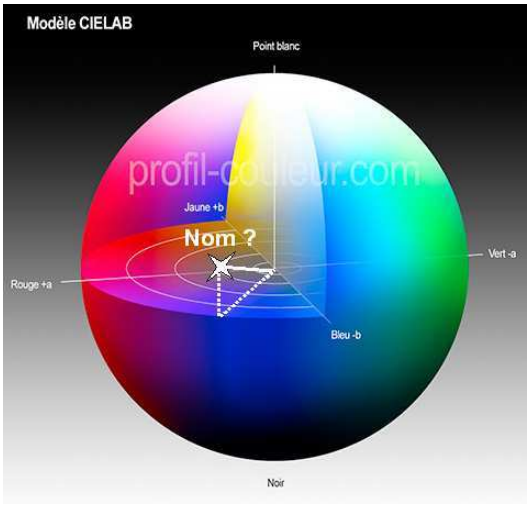



FIGURE 3.12 – Couleur de nom inconnu représentée dans l'espace CIELAB

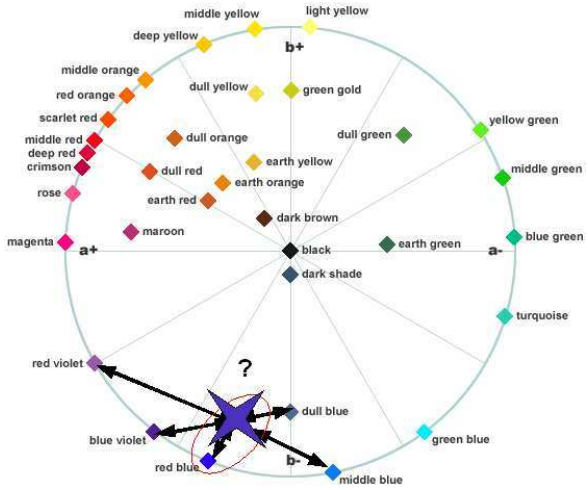



FIGURE 3.13 – Projection de l'espace CIELAB et des couleurs de référence du dictionnaire couleur (seules quelques distances sont représentées)

même couleur de base qui utilisent des adjectifs distinctifs comme Pâle, Foncé, Clair, etc. Ces couleurs semblent assez proches et dire qu'un *"strong pink"* est quasi semblable à un *"vivid pink"* permet une comparaison plus nuancée. Nous avons donc regroupé les couleurs selon leur nom de couleur de base. Ainsi en faisant abstraction des adjectifs introduisant les nuances et en regroupant les entrées du dictionnaire suivant le nom de la couleur de base on obtient 31 noms de couleurs permettant des comparaisons plus grossières. La figure 3.14 montre par exemple deux regroupements possible à partir des couleurs de base "Pink" et "Brown". De cette façon la couleur *"strong pink"* n'est pas différente de la couleur *"vivid pink"* car leurs couleurs de base ont le même nom (voir figure 3.15).

| Pink | |
|---|---|
| Vivid Pink | |
| Strong Pink | |
| Deep Pink | |
| Light Pink | |
| Moderate Pink | |
| Dark Pink | |
| Pale Pink | |
| Grayish Pink | |
| Pinkish White | |
| Pinkish Gray | |

| Brown | |
|---|---|
| Strong Brown | |
| Deep Brown | |
| Light Brown | |
| Moderate Brown | |
| Dark Brown | |
| Light Grayish Brown | |
| Grayish Brown | |
| Dark Grayish Brown | |
| Light Brownish Gray | |
| Brownish Gray | |
| Brownish Black | |

FIGURE 3.14 – Extrait de regroupements des 267 couleurs du ISCC-NBS suivant leur couleur de base

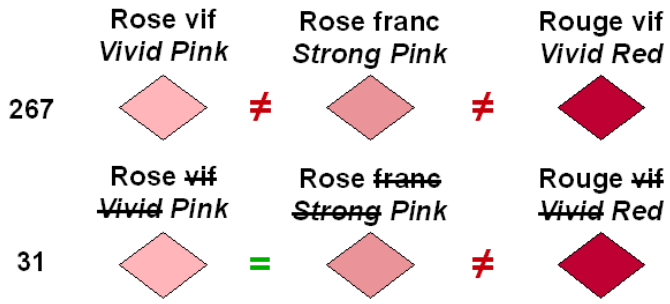



FIGURE 3.15 – Comparaison des noms de couleur avec la méthode pleine (267 noms) et la méthode réduite (31 noms)

Après différents tests réalisés sur des films d'animation (voir annexe C.1) la méthode de comparaison « réduite » aux 31 couleurs de base s'est révélée moins intéressante pour notre objectif de synthétiser l'ensemble du film en minimisant la redondance entre les images. En effet, lorsque l'on compare les images à partir des couleurs de base la comparaison devient trop grossière. Ce phénomène est amplifié par la réduction en blocs des images traitées. Finalement, cette comparaison finit par trop simplifier le film et on perd des passages importants de celui-ci. La méthode « pleine » qui consiste à utiliser le nom complet des couleurs du dictionnaire ISCC-NBS a donc finalement été préférée.

### 3.2.4 Mesure de l'activité

Nous venons de voir que le réducteur, basé sur l'algorithme à accumulation de différences, permet d'obtenir un ensemble d'images clefs liées au changement de contenu dans les images du films. Or, comme nous l'avons vu précédemment, nous partons de l'hypothèse que :

***les zones où les changements de plan ainsi que les changements de contenu sont fréquents correspondent à des passages du film où l'action est élevée.***

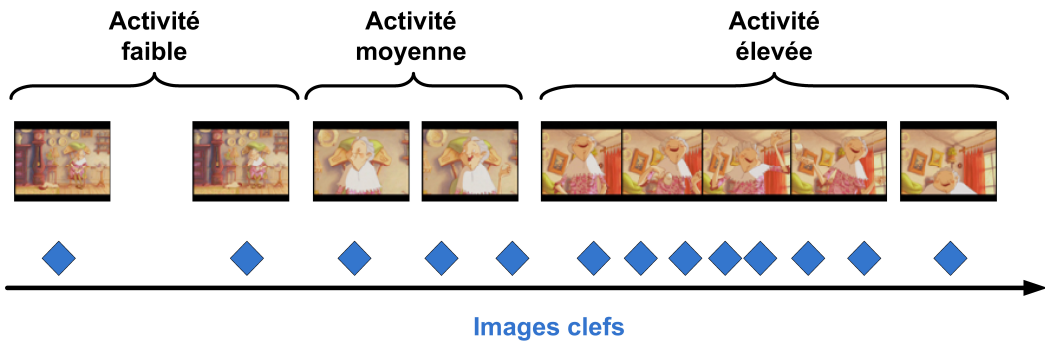



FIGURE 3.16 – Images clefs en sortir du réducteur en fonction de l'activité

L'hypothèse ci dessus se traduit en sortie du réducteur par une densité d'images clefs plus importante dans les passages du film où le contenu évolue rapidement au cours du temps et

où les changements de plan sont fréquents. L'exemple de la figure 3.16 illustre ce principe. Ainsi, nous proposons d'évaluer l'action de deux manières :

- d'abord à travers une description globale qui permet de caractériser l'action contenue dans l'ensemble du film (activité globale).
- ensuite, en construisant une mesure locale qui se présente sous la forme d'une fonction du temps, binaire, indiquant à chaque instant, c'est-à-dire à chaque image de la séquence, s'il y a ou non une action significative (activité locale).

#### 3.2.4.1 Mesure de l'activité globale

Nous proposons de mesurer l'activité globale de la séquence à partir de la distribution des images clefs au cours du temps. Sur la figure 3.16 on voit que plus l'activité est importante, plus la densité d'images clefs est grande et plus le temps séparant ces images clefs est faible. Ainsi, il y a une relation entre l'activité et la fréquence d'apparition des images clefs.

**3.2.4.1.1 Calcul de l'activité** Dans un premier temps nous définissons l'indicateur $\Delta_T(i)$ comme la mesure du temps exprimée en secondes séparant deux images clefs consécutives $i$ et $i+1$ (voir figure 3.17).

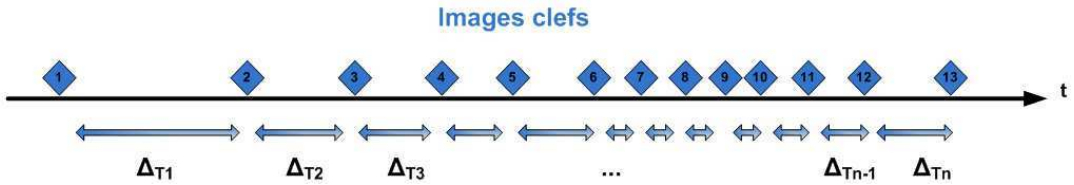



FIGURE 3.17 – Calcul du temps moyen entre les images clefs

Nous calculons ensuite le premier paramètre caractéristique de la séquence qui est l'intervalle de temps moyen séparant deux images clefs successives $E(\Delta_T)$.

$$E(\Delta_T) = \frac{1}{N-1} \sum_{i=1}^{N-1} \Delta_T(i) \tag{3.7}$$

où $N$ est le nombre total d'images clefs extraites de la séquence d'animation et $E(\Delta_T) \in [4 * T_R, +\infty]$. Du fait du mécanisme de compensation de mouvement partiel, la différence d'indice minimum séparant deux images clefs consécutive est de 4. Le temps d'échantillonnage des images dans un film est $T_R = 1/F_R$ où $F_R$ (*"Frame Rate"*) est le nombre d'images par seconde utilisées pour construire le film.

Cependant il est plus commode de caractériser l'activité globale de la séquence vidéo par la fréquence d'apparition des images clefs définie par :

$$F_{moyen} = \frac{1}{E(\Delta_T)} \tag{3.8}$$

avec $F_{moyen} \in [0, F_R/4]$ où $F_R$ (*"Frame Rate"*) est le nombre d'images par seconde utilisées pour construire le film. En général $F_R$ vaut 25 images par seconde.

Or, cette mesure n'est pas normalisée entre 0 et 1. Pour obtenir cette normalisation nous utilisons un coefficient de normalisation correspondant à la valeur maximale possible de $F_{moyen}$ :

$$Activite = \frac{4}{F_R} * F_{moyen} \tag{3.9}$$

Cette caractéristique permet une comparaison **absolue** de l'activité des films. Intuitivement, un film dont la fréquence moyenne est $F_{moyen} = 3$ images clefs par seconde à une activité bien supérieure à un film de fréquence moyenne $F_{moyen} = 0.8$ images clefs par seconde. Cette mesure permet donc une comparaison des films entre eux, ce que le descripteur issu de l'équation 3.6 ne permettait pas.

Comme présenté au début de ce chapitre, cette valeur numérique est transformée en une valeur symbolique par l'utilisation d'ensembles flous. Le concept linguistique *Activité globale de la séquence* est décrit par trois valeurs linguistiques : *"activité faible"*, *"activité moyenne"* et *"activité haute"*. La partition floue $F_{Activite}$ de l'univers de discours, $R_{Activite}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\mu_F$, $\mu_M$ et $\mu_H$ qui constituent le partitionnement de l'univers de discours $R_{Activite}$ noté $L_{Activite}(R_{Activite})$, et est illustrée par la figure E.8. Ces fonctions d'appartenance sont obtenues par expertise du domaine.

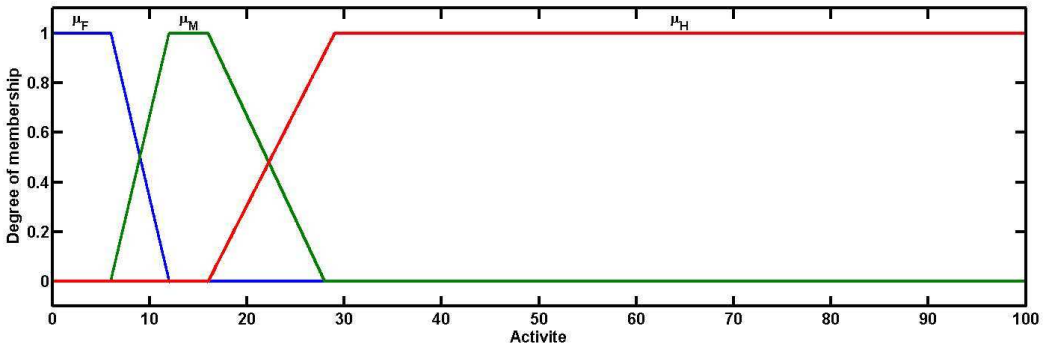



FIGURE 3.18 – La partition floue $F_{Activite}$ de l'univers de discours de la mesure d'activité globale $R_{Activite}$ est déterminée par les fonctions d'appartenance floues : $\mu_F(R_{Activite}) = 1, \forall R_{Activite} \in [0, 6]$, $\mu_M(R_{Activite}) = 1, \forall R_{Activite} \in [12, 16]$ et $\mu_H(R_{Activite}) = 1, \forall R_{Activite} \in [28, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

L'activité globale permet de caractériser l'activité dans la séquence d'animation. Cependant, ce descripteur ne nous renseigne pas sur la distribution de cette activité tout au long de la séquence. Ainsi nous allons calculer un second descripteur, le rythme de la séquence.

**3.2.4.1.2 Calcul du rythme** Une deuxième caractéristique intéressante à extraire est le rythme de la séquence. On cherche à mesurer par l'intermédiaire du rythme si l'activité est présente en continu ou entrecoupée de passages plus calmes. Ainsi, le rythme est lié à l'homogénéité de la séquence vidéo. Numériquement, nous définissons le rythme comme l'écart type de l'indicateur $\Delta_T$.

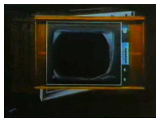
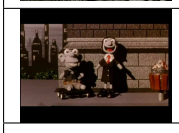
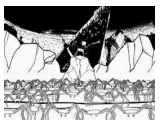
| | Film | Act | Ryt | URL |
|---|---|---|---|---|
|  | Frank Film *Mouris, 1973* | 65% | 7% | http://www.dailymotion.com/video/x700s6_frank-film-frank-mouris-1973_creation |
|  | Dimensions of dialog *Svankmajer, 1983* | 29% | 12% | http://www.dailymotion.com/video/x2gtpo_jan-svankmajer-dimensions-of-dialog_shortfilms |
|  | Harvie Krumpet *Elliot, 2003* | 9% | 12% | http://www.dailymotion.com/video/x8dr7p_harvie-krumpet-vostfr_creation |
|  | La Révolution des crabes *De Pins, 2004* | 8% | 48% | http://www.dailymotion.com/video/x2k6ll_la-revolution-des-crabes_creation |
|  | Au bout du monde *Bronzit, 1998* | 5% | 31% | http://www.dailymotion.com/video/x4v0c4_konstantin-bronzit-au-bout-du-monde_shortfilms |

Figure 3.19 – Valeur de l'activité et du rythme mesurée sur quelques films disponibles via internet.

$$Rythme = \sigma_{\Delta_T} = \sqrt{E[\Delta_T^2] - E[\Delta_T]^2} \tag{3.10}$$

Or, cette définition du rythme n'est pas normalisée entre 0 et 1. Pour obtenir cette normalisation nous utilisons un coefficient de normalisation correspondant à la valeur maximale possible de $\sigma_{\Delta_T}$ :

$$\sigma_{\Delta_T}^{max} = \frac{max(\Delta_T) - min(\Delta_T)}{2} \tag{3.11}$$

Ce coefficient $\sigma_{\Delta_T}^{max}$ majorant $\sigma_{\Delta_T}$, nous définissons ainsi la valeur normalisée du rythme $R_{norm}$ par :

$$R_{norm} = \frac{\sigma_{\Delta_T}}{\sigma_{\Delta_T}^{max}} \tag{3.12}$$

Par analogie avec la musique, l'activité que nous avons définie plus haut correspond au tempo du morceau musicale. C'est la vitesse à laquelle est jouée la rythmique. Alors que le calcul du rythme définit ici donne une image des changements rythmiques du film. Contrairement au descripteur d'activité globale, le descripteur de rythme est moins facile à manipuler. En effet, sur la figure 3.19 les films sont rangés suivant l'activité mesurée et lorsque l'on regarde ces films on « ressent » bien cette hiérarchie. Le descripteur d'activité permet effectivement de comparer les films entre-eux. Cependant comparer les films sur la base du rythme est beaucoup plus difficile en raison du caractère relatif de cette mesure. Cette comparaison est envisageable pour une mesure d'activité identique. En effet, bien que les films *La Révolution des crabes* et *Harvie Krumpet* aient la même mesure d'activité le deuxième film nous apparaît bien plus rythmé que le premier comme le laisse apparaître la mesure du rythme. Cette comparaison sur le rythme n'est cependant plus valide lorsque la mesure d'activité est différente (par exemple entre *Dimensions of dialog* et *Harvie Krumpet* voir figure 3.19).

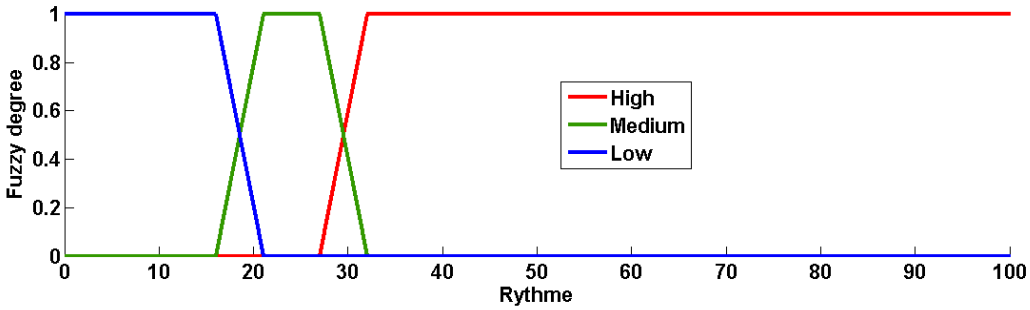



FIGURE 3.20 – La partition floue $F_{Rythme}$ de l'univers de discours de la mesure du rythme $R_{norm}$ est déterminée par les fonctions d'appartenance floues : $\mu_{Lent}(R_{norm}) = 1, \forall R_{norm} \in [0, 16]$ (bleu), $\mu_{Moyen}(R_{norm}) = 1, \forall R_{norm} \in [21, 27]$ (vert) et $\mu_{Rapide}(R_{norm}) = 1, \forall R_{norm} \in [32, 100]$ (rouge), (l'axe des ordonnées correspond au degré d'appartenance).

De plus, comme pour le descripteur d'activité globale, cette valeur numérique est transformée en une valeur symbolique par l'utilisation d'ensembles flous. Le concept linguistique *Rythme de la séquence* est décrit par trois valeurs linguistiques : *"rythme lent"*, *"rythme moyen"* et *"rythme rapide"*. La signification floue de chaque symbole est illustrée par sa fonction d'appartenance floue (voir la figure 3.20). Les partitions floues ont été déterminées par expertise du domaine.

### 3.2.4.2 Mesure de l'activité locale

Nous cherchons à retrouver les passages de la vidéo où l'activité est significative par rapport à l'ensemble de la séquence. Nous adoptons la même démarche que dans [Ionescu, 2007].

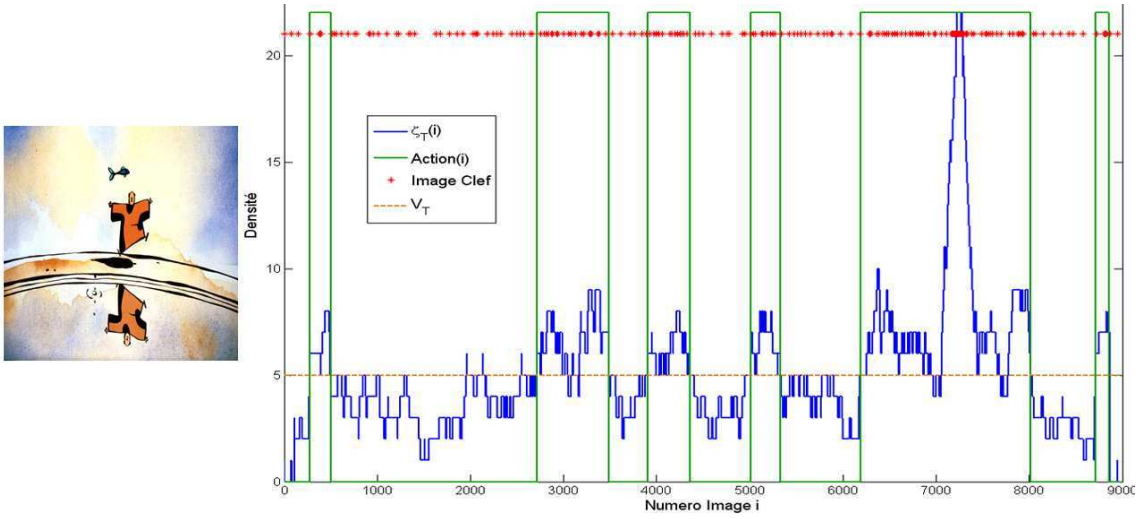



FIGURE 3.21 – Exemple de la fonction **Action(i)** pour le film "Le moine et le poisson"

Plus précisément, la mesure est effectuée de la manière suivante :

- dans un premier temps, nous définissons un indicateur de base relié à la structure temporelle de la séquence. Cet indicateur, noté $\zeta_T(i)$, où $i$ désigne le numéro de l'image, représente le nombre d'images clefs dans une plage de durée $T$ (typiquement $T = 5$s).
- dans un second temps nous définissons les *segments d'action* définis par seuillage de la fonction $\zeta_T(i)$.

$$Action(i) = \begin{cases} 1 & \text{si } \zeta_T(i) > \bar{v}_T \\ 0 & \text{sinon} \end{cases} \tag{3.13}$$

  Ce qui signifie que l'action est considérée comme significative si le nombre d'images clefs par durée de temps $T$ (ou densité) est supérieur à la moyenne $\bar{v}_T$ de $\zeta_T(i)$ calculée sur l'ensemble de la séquence.
- enfin, nous procédons à quelques post-traitements (ouvertures et fermetures morphologiques) sur la fonction "segments d'action" (élimination des segments trop courts et fusion des segments très proches) obtenant ainsi la fonction *"Action(i)"*.

La figure 3.21 donne une illustration de la mesure de l'action locale pour le film "Le moine et le poisson". Notons enfin que cette mesure de l'action locale présente un caractère relatif puisque le seuil de binarisation est dépendant du contenu de chaque film.

### 3.2.5 Le condenseur

Comme on l'a vu précédemment, l'algorithme à accumulation est un pré-traitement pour l'étage du condenseur. En effet, cet étage permet la création d'un condensat (nécessaire au calcul des caractéristiques couleurs) en ne sélectionnant qu'un nombre limité d'images représentatives de la séquence entière.

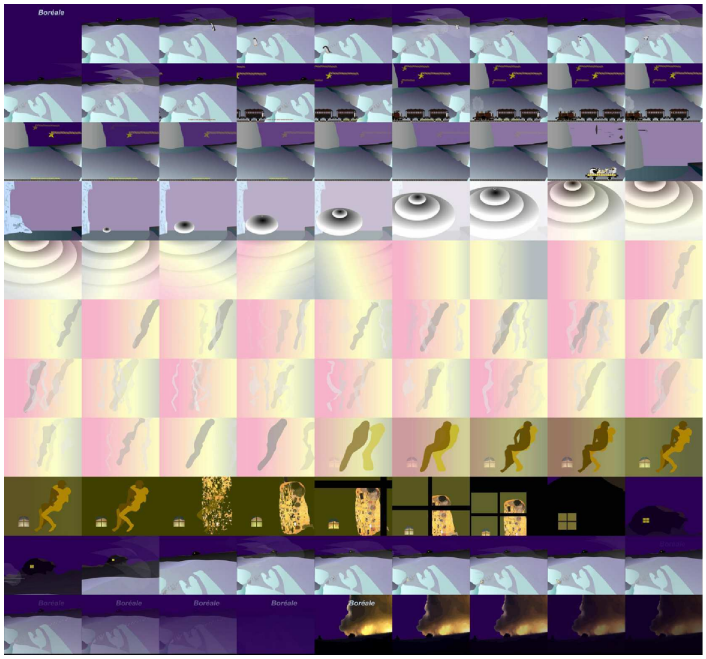


FIGURE 3.22 – Ensemble des images clefs à la sortie de l'algorithme à accumulation sur un film d'animation (*Boréale, (3197 images)*)

En moyenne l'algorithme à accumulation réduit la longueur du film à 4.2% [1] de sa longueur originale. Bien que cette étape réduise considérablement la redondance de l'information il reste encore beaucoup trop d'images et l'information restante n'est pas vraiment synthétisée (voir figure 3.22). De plus, on retrouve encore beaucoup de redondance si le film analysé est construit avec un plan qui revient plusieurs fois dans le temps.

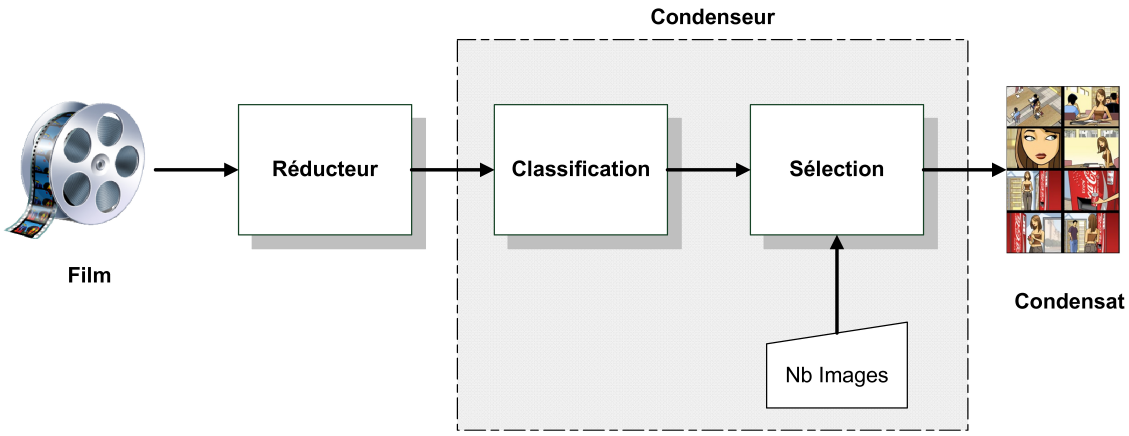



FIGURE 3.23 – Chaine de traitements du condenseur

Nous devons sélectionner un certain nombre d'images dans l'ensemble des images issues du réducteur. Ces images doivent donc être différentes entre elles afin d'obtenir une réduction du nombre d'images composant le condensat la plus efficace possible. Pour réaliser cette étape nous devons regrouper les images par similarité visuelle. C'est l'étape de classification (voir figure 3.23). Enfin, l'étape de sélection (voir figure 3.23) permet de sélectionner une image représentative de chacun des regroupements (clusters) pour constituer le condensat.

#### 3.2.5.1 La Classification Ascendante Hiérarchique

Classifier, c'est regrouper entre eux des objets similaires selon un critère prédéfini. Les techniques de classification visent à répartir $n$ individus, caractérisés par un ensemble de variables $X_1, X_2, \ldots, X_p$ en $m$ sous-groupes les plus homogènes possibles. On distingue deux grandes familles de techniques de classification :

- **La classification non hiérarchique ou partitionnement**, aboutissant à la dissociation de tous les individus en $m$ classes d'équivalence où le nombre $m$ de classes est fixé *a priori*.
- **La classification hiérarchique** : à un niveau $l$ donné, deux individus peuvent être regroupés dans un même groupe, alors qu'à un niveau $l+1$, ils seront dissociés et appartiendront à deux sous-groupes différents.

La distinction entre ces deux familles vient du choix du nombre de classes. Dans notre cas ce nombre de classes n'est pas connu *a priori* c'est pourquoi on préférera une Classification Ascendante Hiérarchique (CAH). De plus, ce qui nous intéresse ce n'est pas la hiérarchie, mais une **typologie**, c'est-à-dire une partition de l'ensemble des données en clusters. Pour obtenir l'ensemble des clusters finaux il est nécessaire de définir et de passer par les étapes suivantes :

1. tests réalisés sur une base très diversifiée de 107 films d'animation.

1. Construire la hiérarchie
   Étant donné un cluster, quelle est la meilleure manière de le scinder en deux "enfants" ? Ou bien, à l'inverse, comment choisir deux clusters dans le but de les fusionner en un unique cluster parent ? Ces deux questions donnent naissance, respectivement, aux Hiérarchies Descendantes et Ascendantes.
2. Choisir une typologie dans la hiérarchie
   Étant donnée une hiérarchie, quelle section de l'arbre doit être retenue comme typologie finale ?

La **CAH** procède par fusions successives de clusters déjà existants. A chaque étape, les deux clusters qui vont être fusionnés sont ceux dont la "distance" est la plus faible. Il faut donc définir la "distance" entre deux individus puis la "distance" entre deux groupes d'individus. Initialement la CAH considère toutes les observations comme étant des clusters ne contenant qu'une seule observation (singleton). La première étape consiste à fusionner dans un cluster les deux observations les plus proches (au sens de la distance entre individus choisie), puis de continuer, en fusionnant à chaque étape les deux clusters les plus proches. Le processus s'arrête quand les deux clusters restant fusionnent dans l'unique cluster contenant toutes les observations.

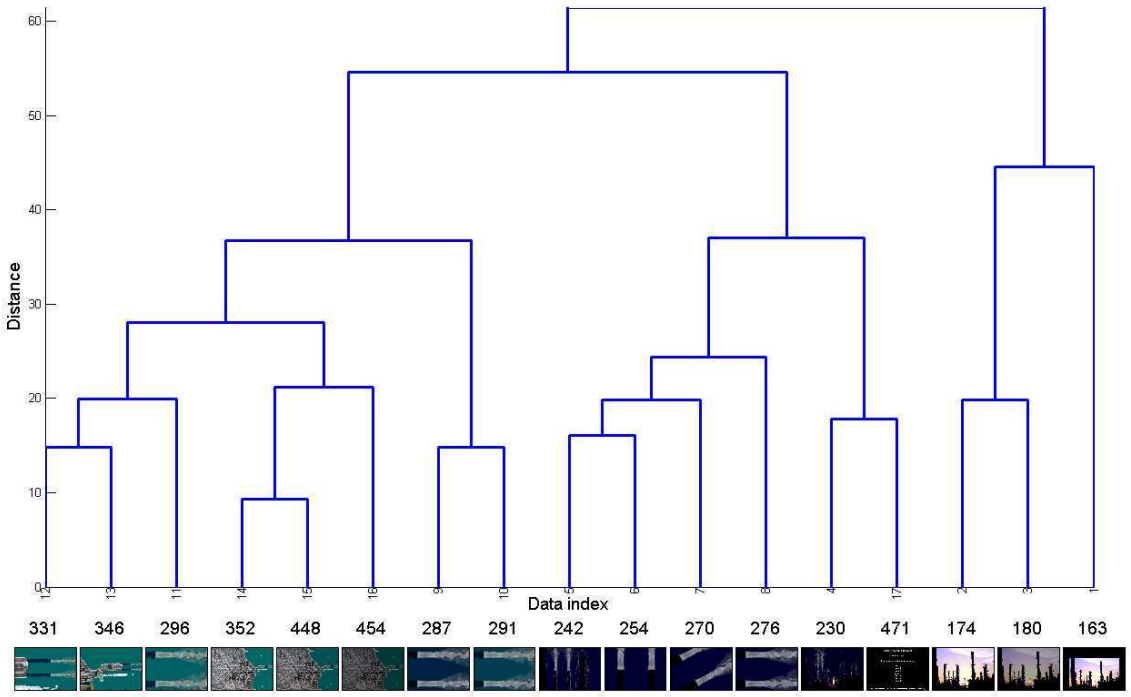



FIGURE 3.24 – Dendogramme et les images correspondantes *(mesure de la différence en ordonnée et index de l'image dans la vidéo en abscisse).*

La représentation la plus communément utilisée pour tracer l'arbre de la hiérarchie ainsi obtenu est le dendogramme (voir figure 3.24). Une règle généralement utilisée pour obtenir les typologies qui ont le plus de chance d'être significatives est de tracer une ligne horizontale

en travers du dendogramme, et de ne retenir dans la typologie que les clusters terminaux qui sont juste au-dessus de cette ligne [Ennaji *et al.*, 2003]. En changeant la hauteur de la ligne, on change le nombre de clusters retenus, et on dispose ainsi d'un moyen simple pour faire varier la granularité de la typologie finale. Ainsi, si l'on connait le nombre de clusters ou plus particulièrement le nombre d'images composant le condensat (en considérant que l'on extrait une image par cluster), on peut construire facilement un ensemble d'images statiques fortement représentatives de la séquence. Mais cette connaissance, *a priori* du nombre de clusters est rarement à la disposition de l'utilisateur. Plusieurs méthodes ont été proposées pour déterminer le point de coupure et ainsi trouver automatiquement le nombre de clusters [Calinski et Harabasz, 1974, Milligan et Cooper, 1985, Krzanowski et Lai, 1985, Ennaji *et al.*, 2003].

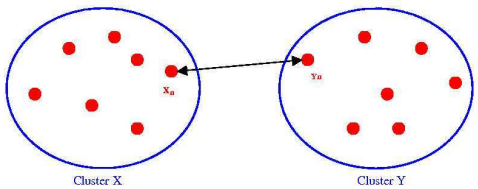
La définition d'une distance entre clusters demande la définition préalable d'une distance entre les individus à classifier (dans notre cas les individus sont des images). De nombreuses distances ont été utilisées afin de juger de la similarité entre les images (voir annexe C.2). Finalement la méthode par **décomposition en bloc** est un bon compromis entre vitesse et performance et est définie comme suit :

$$\begin{aligned} d(f_i, f_j) &= \frac{1}{N^2}\sum_{k=1}^{N^2}(DeltaE76(F_i(k), F_j(k))) \\ DeltaE76(P_x(L,a,b), P_y(L,a,b)) &= \sqrt{(L_x - L_y)^2 + (a_x^* - a_y^*)^2 + (b_x^* - b_y^*)^2} \end{aligned} \quad (3.14)$$

où $f_i$ et $f_j$ désignent deux images. Chacune de ces images est transformée, par découpage en blocs, en une matrice réduite ($F_i$ et $F_j$) de taille N*N (comme expliquée dans 3.2.3.1.2 où est extrait pour chaque cellule la valeur médiane vectorielle des pixels composant le bloc). Cette valeur médiane est représentée par un triplet de valeur (L,a*,b*) dans l'espace CIE-Lab. $N$ est fixé pour obtenir des blocs d'une cinquantaine de pixels (par exemple sur une image de 800*600 pixels on fixe $N = 97$). On calcule les $N^2$ distances, basées sur la formule de différence de couleur (DeltaE76) du système colorimétrique ClE1976Lab où la distance entre les composantes Lab des deux points ($P_x$ et $P_y$) de cet espace est basée sur la distance euclidienne. Ensuite la moyenne de ces $N^2$ distances sert de métrique pour la comparaison des deux images (équation 3.14).

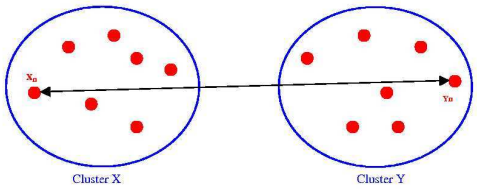
A partir de cette distance entre individus il reste à déterminer la distance $D(X,Y)$ entre deux groupes d'individus ou clusters $X$ et $Y$. Généralement les distances utilisées sont le minimum, le maximum ou la moyenne pondérée des distances entre les singletons ($x_n$) qui constituent les deux clusters. Les figures ci dessous illustrent ces distances :

– **Saut minimum "Single linkage" :**




$$D(X,Y) = min(d(x,y)) \text{ where } x \in X, y \in Y \quad (3.15)$$

– **Saut maximum “Complete linkage" :**




$$D(X,Y) = max(d(x,y)) \text{ where } x \in X, y \in Y \tag{3.16}$$

– **Saut moyen “Average linkage" :**

$$D(X,Y) = \frac{1}{\Omega(X) * \Omega(Y)} \sum_{i=1}^{\Omega(X)} \sum_{j=1}^{\Omega(Y)} d(x_i, y_j) \text{ where } x_i \in X, y_j \in Y \tag{3.17}$$

Après différents tests nous avons retenu la méthode de clustering *Complete linkage* qui donne de bons résultats (voir annexe C.2 pour les différents tests). L'ensemble des paramètres de l'algorithme de classification étant fixés, nous obtenons un dendogramme (un exemple d'un tel dendogramme est représenté sur la figure 3.24). Cette représentation fait apparaître les hiérarchies de clusters et les images correspondantes. Finalement, une sélection des images dans chacun des clusters est nécessaire pour construire le condensat d'images statiques du film d'animation.

#### 3.2.5.2 La sélection des images

La CAH permet d'obtenir $N_{clusters}$ clusters d'images similaires. Il faut donc extraire pour chaque cluster une image représentative. On propose de prendre l'image médiane $I_{med}(C)$ du cluster $C$, c'est-à-dire l'image dont la distance cumulée aux autres images du cluster est la plus faible. En traitement d'images à plusieurs composantes (image couleur par exemple), la formulation classique des filtres médian vectoriel consiste à calculer la sortie du filtre comme le vecteur qui minimise la somme des écarts cumulés à tous les autres vecteurs de la fenêtre de filtrage. On obtient ainsi le vecteur “le plus représentatif” de la fenêtre. Les écarts entre images sont fournis par la mesure de distance définie par l'équation 3.14. Nous introduisons la distance cumulée de l'image $i$ aux autres images du cluster comme suit :

$$D_{cum}(f_i) = \sum_{f_j \in C, f_j \neq f_i} d(f_i, f_j) \tag{3.18}$$

où $D_{cum}(f_i)$ est la distance cumulée de l'image $f_i$ aux autres images du cluster $C$, $d(f_i, f_j)$ est la distance définie précédemment (équation 3.14). Enfin, l'image représentative du cluster est l'image $f_i$ dont la distance cumulée $D_{cum}(f_i)$ est la plus petite parmi les distances cumulées du cluster $C$ (equation 3.19).

$$I_{med}(C) = \arg\min_{f_j \in C} D_{cum}(f_j) \tag{3.19}$$

L'image médiane peut être choisie comme l'image représentative de l'ensemble des images du cluster. La figure 3.25 donne un exemple d'une telle sélection. Ainsi les images 331, 448 et 254 sont les images représentatives de leur cluster respectif (délimité par des pointillés verts). Dans le cas particulier des clusters à 2 éléments nous avons fait le choix arbitraire de sélectionner la deuxième image du cluster (par exemple les images 291, 471 et 180 sur la figure 3.25). Lorsque le cluster ne contient qu'une seule image le choix de son image représentative est trivial.

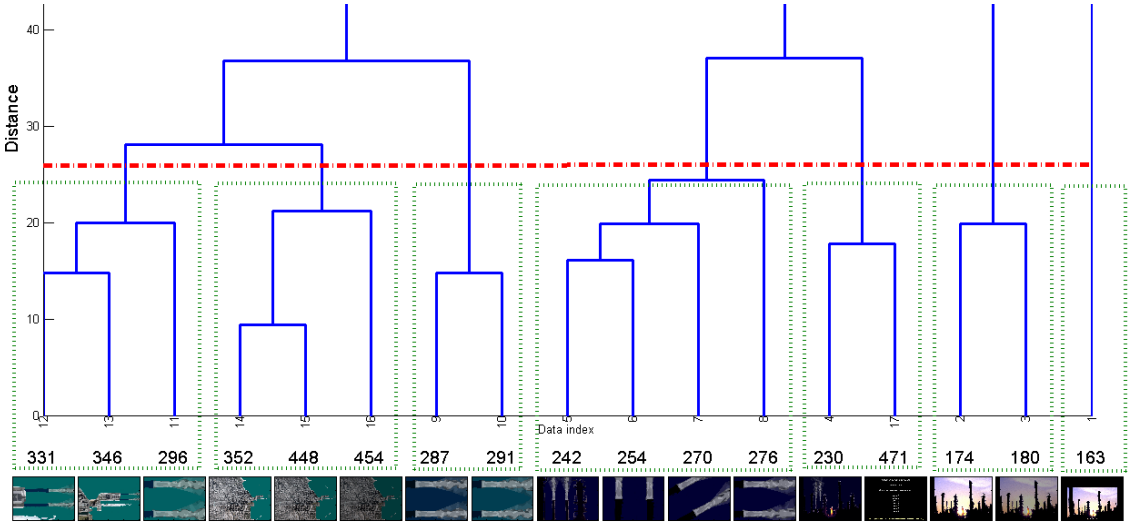



FIGURE 3.25 – Coupure du dendogramme pour obtenir 7 clusters. *(Le seuil de coupure du dendogramme (distance $\simeq 27$) est en rouge. Les clusters ainsi obtenus sont en vert)*

Après l'étape de sélection nous obtenons le condensat, sorte de « résumé » de la séquence vidéo composé d'un ensemble de $N_{clusters}$ images représentatives. Rappelons ici que notre objectif initial (voir la figure 3.6) est d'obtenir la caractérisation couleur de la séquence vidéo à partir de l'analyse des histogrammes couleurs. Or, on l'a vu précédemment, cette caractérisation globale de la séquence s'appuie sur le calcul d'un histogramme global pondéré (voir l'équation 3.2). Cet histogramme est en réalité la somme pondérée de tous les histogrammes moyens calculés pour chacune des images représentatives de la séquence, où cette pondération $\omega_i$ représente le pourcentage d'images de la séquence vidéo appartenant au plan $i$. Elle dépend donc de la longueur du plan vidéo considéré. Dans notre approche l'utilisation des plans n'est plus considérée et par conséquent leur longueur n'est pas disponible. Cependant, pour conserver ce mécanisme de pondération et donc accorder une importance différente aux histogrammes calculés à partir des images du condensat, nous proposons une nouvelle définition de cette pondération comme étant la longueur du cluster $C_i$ auquel appartient l'image $i$. Nous définissons la longueur d'un cluster comme étant la somme des longueurs des sous-séquences d'images composant ce cluster. Une sous-séquence d'images clefs est un ensemble ordonné (suivant les indices) d'au moins deux images clefs appartenant au même cluster. Sa longueur est définie comme le nombre d'images (ou différence d'indice) entre la première et la dernière image de la sous-séquence.

Les sous-séquences d'images dans un même cluster sont généralement dues à des plans

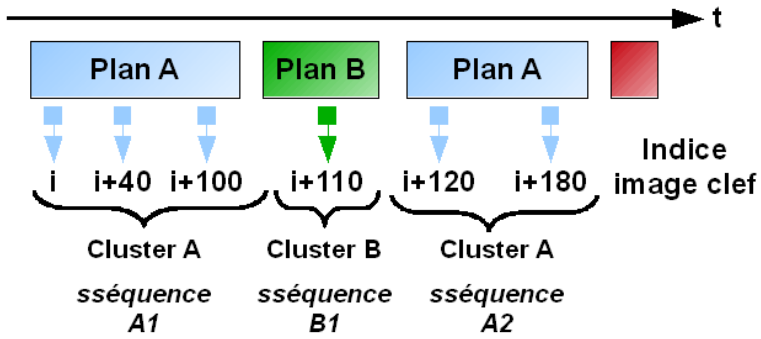



FIGURE 3.26 – Illustration d'une sous-séquence dans un cluster

contenant le même contenu mais revenant à différent moment dans le film comme sur la figure 3.26. Puisque ces plans ont un même contenu les images clefs extraites de ces passages du film seront regroupées au sein d'un même cluster par l'étape de sélection. Ainsi dans l'exemple de la figure 3.26 le cluster $A$ contient 2 sous-séquences ($A1 \in [i, i+40, i+100]$ et $A2 \in [i+120, i+180]$ respectivement issues des deux plans $A$). La longueur d'une sous-séquence d'images est définie comme la différence d'index entre la dernière et la première image de la sous-séquence. Dans l'exemple présenté par la figure 3.26 la longueur de la sous-séquence $A1$ vaut $L_{A_1} = 100$ et la longueur de la sous-séquence $A2$ vaut $L_{A_2} = 60$. Finalement la longueur du cluster $A$ est la somme des longueurs de ses sous-séquences soit $L_A = 160$.

| N° Image Clef | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster | c7 | c6 | c6 | c5 | c4 | c4 | c4 | c4 | c3 | c3 | c1 | c1 | c1 | c2 | c2 | c2 | c5 |
| N° index seq | 163 | 174 | 180 | 230 | 242 | 254 | 270 | 279 | 287 | 291 | 296 | 331 | 346 | 352 | 448 | 454 | 471 |
| Lseq | | 6 | | | 37 | | | | 4 | | 50 | | | 102 | | | |

| Cluster | c1 | c2 | c3 | c4 | c5 | c6 | c7 |
|---|---|---|---|---|---|---|---|
| Lcluster | 50 | 102 | 4 | 37 | 1 | 6 | 1 |

FIGURE 3.27 – Calcul des longueurs des sous-séquences composant les clusters *(Les sous-séquences sont en couleur)*

A partir du dendogramme de la figure 3.25 on construit le tableau de la figure 3.27. Ce tableau représente les images clefs ordonnées suivant leur numéro d'apparition ($N^\circ$ *Image Clef*). Y figure également le numéro du cluster auquel appartient l'image (*Cluster*), ainsi que son index dans la séquence vidéo ($N^\circ$ *index seq*).
A partir de ce tableau il est facile de retrouver les sous-séquences d'images. Ce sont les images dont les indices $N^\circ$ *index seq* se suivent et qui ont le même numéro de cluster (cellules en couleur sur la figure 3.27). La longueur d'une sous-séquence d'images est définie comme la différence d'index entre la dernière et la première image de la sous-séquence. Par exemple, sur la figure 3.27, le cluster $C_1$ est constitué de l'unique sous-séquence d'images $S_{C1} = [296, 331, 346]$ de longueur $L_{seq} = 346 - 296 = 50$ images. De plus, le cluster $C_5$ constitue un cas particulier. En effet, il est constitué de deux images très sombres correspondant au début (image d'indice 230) et à la fin du film (image d'indice 471). Ce cluster, dont l'image représentative est arbitrairement choisie comme la deuxième image (image d'indice 471), ne contient pas de sous-séquence (ou d'images consécutives). En effet les deux images

de ce cluster ne constituent pas une sous-séquence puisque d'autres images de clusters différents les séparent. Par conséquent la longueur de ce cluster est égale à la longueur d'une image.

Finalement pour chaque cluster $i$ nous définissons un poids $\omega_i$ :

$$\omega_i = \frac{L_{cluster}(i)}{\sum_{j=1}^{N_{clusters}} L_{cluster}(j)} \tag{3.20}$$

$$L_{cluster}(j) = \sum_{k=1}^{N_{sseq}^{j}} L_{sseq}(k) \tag{3.21}$$

où $N_{sseq}^{j}$ est le nombre de sous-séquences d'images dans le cluster $j$, $L_{sseq}(k)$ est la longueur de la sous-séquence $k$ du cluster $j$ et $N_{clusters}$ est le nombre de clusters dans la séquence vidéo.

$$h_{seq}(c) = \sum_{i=1}^{N_{clusters}} h_i(c) \cdot \omega_i \tag{3.22}$$

Pour chaque cluster $i$ nous calculons l'histogramme des couleurs réduites $h_i(c)$ ($c$ l'indice de la couleur) de l'image $I_{med}(i)$ comme présenté au début de ce chapitre (voir équation 3.1). L'histogramme global pondéré de la séquence est ensuite calculé (voir équation 3.22) comme la somme pondérée par $\omega_i$ de l'histogramme de chacun des clusters $h_i(c)$ ($i = 1, ..., N_{clusters}$ et $N_{clusters}$ le nombre total de clusters).

Le processus de création du condensat ("condenseur") permet de fournir un ensemble d'images représentatives de la séquence vidéo afin de calculer un histogramme global pondéré des couleurs permettant ainsi le calcul des descripteurs symboliques décrits précédemment. Toutefois, ce nombre d'images (ou de clusters) doit être fixé par l'utilisateur (voir figure 3.23). Nous allons voir comment fixer ce paramètre.

**3.2.5.2.1 Le choix du pourcentage d'images utilisées** La qualité de la représentation de la distribution globale des couleurs par l'histogramme global pondéré est liée à la valeur du nombre d'images retenues pour construire le condensat. Afin de déterminer la valeur optimale de ce paramètre ($N_{clusters}$) nous avons effectué l'étude suivante :

Nous partons de la représentation la plus fidèle de la distribution globale des couleurs dans la séquence vidéo obtenue en utilisant toutes les images de la séquence. L'histogramme global obtenu dans cette situation, $\widetilde{h_{seq}}(c) = h_i(c)|_{\forall i \in seq}$, est utilisé comme *référence* pour mesurer la qualité de la représentation de la distribution globale des couleurs $h_{seq}(c)$ obtenue pour différentes valeurs de $N_{clusters}$.

Pour faciliter l'étude nous définissons le paramètre $N_{\%}$ comme étant le pourcentage d'images clefs conservées pour composer le condensat et in fine, pour calculer l'histogramme global des couleurs $h_{seq}(c)$.

$$N_{\%} = 100 * \frac{N_{clusters}}{N_{images\ clefs}} \tag{3.23}$$

Pour trouver la valeur optimale de $N_{\%}$ nous avons calculé plusieurs histogrammes globaux pondérés pour différentes valeurs de $N_{\%}$. Nous les avons ensuite comparés à la référence $\widetilde{h_{seq}}(c)$. Comme mesure de similarité nous avons utilisé la distance euclidienne, $d_E(h1, h2)$ entre l'histogramme global pondéré obtenu pour un pourcentage $N_{\%}$ d'images $(h_{seq}(c)|_{N_{\%}})$ et l'histogramme de référence $\widetilde{h_{seq}}(c)$ calculé à partir de toutes les images de la séquence vidéo :

$$d_E(\widetilde{h_{seq}}, h_{seq}|_{N_{\%}}) = \sqrt{\sum_{c=1}^{216} (\widetilde{h_{seq}}(c) - h_{seq}(c)|_{N_{\%}})^2} \tag{3.24}$$

où $c$ est l'indice des couleurs de la palette "Webmaster" de 216 couleurs.

Nous avons également calculé les histogrammes globaux non pondérés pour différentes valeurs de $N_{\%}$. Puis nous les avons comparés à la référence $\widetilde{h_{seq}}(c)$. Un histogramme global non pondéré $h_{seq*}(c)$ est défini comme la moyenne arithmétique de chacun des histogrammes calculés à partir des images du condensat (voir equation 3.25).

$$h_{seq*}(c) = \frac{1}{N_{clusters}} \sum_{i=1}^{N_{clusters}} h_i(c) \tag{3.25}$$

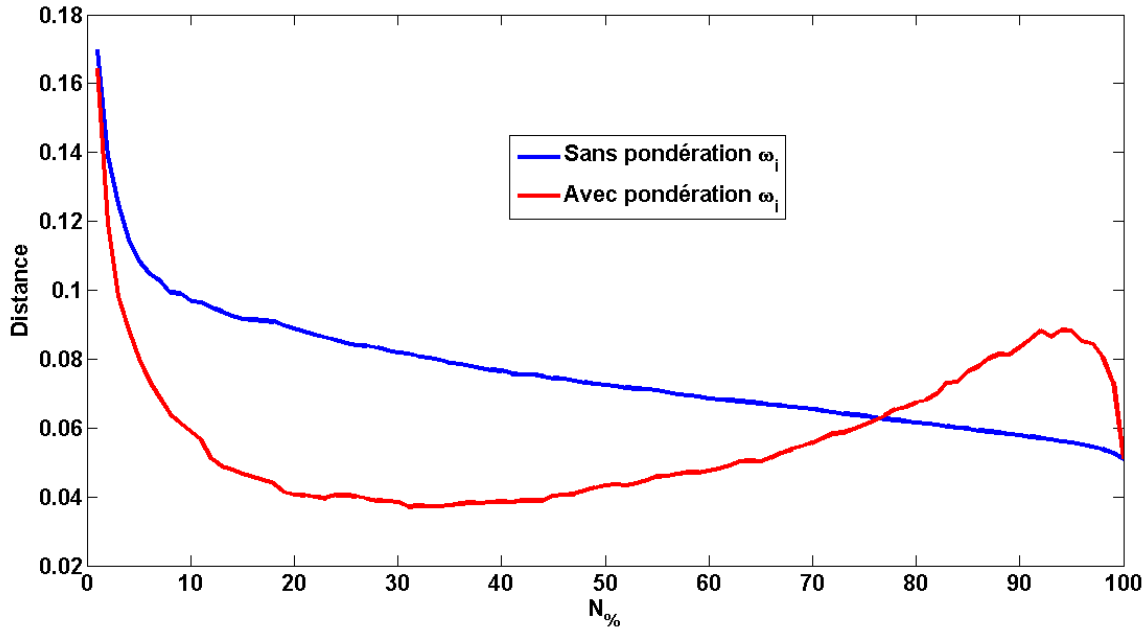



FIGURE 3.28 – Influence du paramètre $N_{\%}$ sur le calcul de l'histogramme global

La figure 3.28 présente les résultats des tests effectués sur une base de 100 films d'animation. Les distances sont normalisées entre 0 et 1 en utilisant un coefficient de normalisation correspondant à la valeur maximale possible de la distance $d_E$ soit $\frac{1}{\sqrt{2}}$. Une distance de 0

signifie que l'histogramme global calculé à partir du condensat est parfaitement identique à l'histogramme de référence $\widetilde{h_{seq}}(c)$. Au contraire une distance de 1 signifie que les histogrammes sont totalement différents. On remarque d'après cette figure que l'utilisation d'un histogramme global pondéré (courbe rouge) est un bon choix par rapport à l'utilisation d'un histogramme global non pondéré (courbe bleue). En effet, on remarque que d'une façon générale l'erreur commise est moins importante dans le cas de la pondération. Comme on pouvait s'y attendre l'erreur diminue lorsque l'on augmente le nombre d'images dans le condensat. Cependant, dans le cas de la pondération, on remarque que l'erreur augmente à partir de $N_{\%} = 50\%$, ceci s'explique par le fait que l'on diminue l'influence des « gros » clusters (correspondant à des passages longs du film où l'activité est faible) en les morcelant et en donnant du même coup plus d'importance à des images atypiques issues de passages courts du film. Finalement, la valeur optimale du nombre d'images $N_{\%}$ se situe aux alentours de 30%. A cette valeur l'erreur commise (au sens de la distance entre histogrammes) est inférieure à 4%. La création du condensat nécessaire au calcul des caractéristiques couleur du film nécessite de choisir seulement 30% des images clefs issues du réducteur pour composer le condensat. Le calcul des caractéristiques couleur est donc accéléré tout en commettant en moyenne une erreur inférieure à 4%.

## 3.3 Conclusion

Dans ce chapitre nous avons vu comment extraire de l'information à partir des images. Cette caractérisation d'un niveau sémantique relativement faible, passe par la caractérisation des couleurs utilisées dans la séquence vidéo ainsi que la caractérisation de l'activité et du rythme. Ces caractérisations s'appuient sur une analyse du changement de contenu s'opérant dans la séquence d'images et utilise un algorithme à accumulation de différences.

Afin d'obtenir une caractérisation plus riche de la séquence d'animation, l'apport de connaissances et d'informations sémantiques externes est nécessaire. Cet apport d'information externe passe dans notre approche par l'analyse du texte contenu dans les péri-textes et notamment les synopsis des films d'animation. C'est ce que nous allons voir dans le chapitre suivant.

Chapitre 4

# Extraction d'information à partir des textes

**Résumé :** Nous avons vu dans les précédents chapitres que l'apport de connaissances et d'informations *a priori* est nécessaire, tant à des fins de caractérisation de séquences d'animation que pour l'extraction de caractéristiques à partir de l'image. Cet apport d'informations peut se faire par l'analyse des synopsis des fiches d'inscription des films d'animation enregistrés auprès du CITIA. L'approche présentée dans ce chapitre consiste à analyser automatiquement ces textes afin d'en extraire une information utile pour caractériser la séquence d'animation. Deux grandes approches sont envisagées ici : d'abord une analyse statistique du corpus qui permet d'appréhender les textes et leurs vocabulaires et qui permet de faire de l'analyse thématique pour la recherche d'atmosphère liée à un genre, ici le Drame. Ensuite, une analyse par extraction d'information permet via la structure syntaxique de la phrase, d'extraire des informations pertinentes modélisées sous la forme d'un *scénario actanciel*.

Nous avons vu dans les précédents chapitres que l'apport de connaissances et d'informations *a priori* est nécessaire, tant à des fins de caractérisation de séquences d'animation que pour l'extraction de caractéristiques à partir de l'image. Cet apport d'information peut se faire par l'analyse des synopsis des fiches d'inscription des films d'animation enregistrés auprès du CITIA lors du festival (voir la figure 2.4). L'approche présentée dans ce chapitre consiste à analyser automatiquement ces textes afin d'en extraire une information utile pour caractériser la séquence d'animation. Ainsi nous aborderons dans l'état de l'art les méthodes généralement mises en œuvre pour caractériser des corpus textuels. Nous présenterons ensuite les résultats d'une analyse statistique sur le corpus des synopsis français qui nous permettra de repérer un ensemble de caractéristiques de ces textes. Cette étape préalable de caractérisation de l'ensemble des synopsis nous permet par la suite de caractériser chacun des synopsis afin d'en extraire une information pertinente et structurée grâce à l'étape d'extraction d'information. Cette étape basée sur un ensemble d'analyses du lexique et de la structure du texte permet d'isoler les actions mises en jeu dans le film. Finalement une deuxième caractérisation du synopsis sera abordée à travers l'analyse thématique. Cette analyse nous permet le repérage des atmosphères dégagées par le texte comme par exemple celle du Drame.

## 4.1 Bref état de l'art

A l'heure actuelle, on estime que la quantité d'information stockée dans le monde double tous les vingt mois (que ce soit dans les pages web, dans les librairies en ligne ou dans les serveurs de courrier électronique). Il est donc indispensable de créer des outils permettant d'exploiter ces informations. Le domaine de l'analyse automatique de texte s'est développé pour répondre à cette volonté de gestion par le contenu des sources volumineuses de textes. Bien que les techniques se soient largement développées ces dernières années, l'analyse de l'ensemble des informations présentes dans un texte est un processus très complexe et reste encore à l'heure actuelle limitée à un domaine précis ou restreint à une compréhension basique du sens. Un grand nombre de disciplines et de domaines de recherche, à la croisée des chemins entre la linguistique, les mathématiques et l'informatique tentent de répondre à des objectifs bien définis et passent par les questions : Quel type de texte analyse-t-on ? Pour répondre à quelles questions ? Désire-t-on étudier le vocabulaire d'un texte en vue d'une analyse du style ? Cherche-t-on à repérer les contenus ? etc.

### 4.1.1 La statistique textuelle

Le texte constitue un passage obligé dans des disciplines très différentes (recherche documentaire, extraction d'informations, texte mining, etc.) dont l'analyse fait intervenir, à des degrés différents, la linguistique et les technologies de l'informatique.

La linguistique est l'étude du langage humain. Elle regroupe un certain nombre d'écoles qui n'abordent pas cet objet d'étude (le langage) du même point de vue. La linguistique structurale étudie la langue comme un système doté d'une structure décomposable. Elle est fondée dès 1910 par Ferdinand de Saussure [Saussure *et al.*, 1922] et reste le courant dominant jusque dans les années 70. La linguistique énonciative, qui hérite de la linguistique structurale, étudie l'acte de produire un énoncé et pas simplement l'énoncé lui-même. D'autres linguistiques sont associées à une discipline particulière (sociologie, ethnologie, psychologie, neurologie, etc). De plus, dans une vision structurelle du langage, on distingue plusieurs domaines qui étudient des faits de langue de natures différentes :

- la **phonétique** : étudie les différents phones ou sons produits par l'appareil phonatoire humain.

- la **phonologie** : étudie comment sont agencés les phonèmes (36 en français) d'une langue pour former des mots. Il ne faut pas confondre la phonologie avec la phonétique qui, elle, s'intéresse aux sons eux-mêmes, indépendamment de leur fonctionnement les uns avec les autres.

- la **morphologie** : étudie la façon dont se combinent les morphèmes ou éléments variables dans les mots pour former des lemmes. Par exemple la dérivation lexicale du préfixe *"re"* et du radical *"partir"* donne *"repartir"*. Dans l'énoncé *"Aller au marché"*, *"au"* est un amalgame des morphèmes *"à"* et *"le"*.

- la **lexicologie** : étudie les lemmes (ou vocabulaires dans les usages courants) composant le lexique d'une langue (environ 60 000, en français liste non exhaustive).

- la **syntaxe** : étudie les relations des mots composant la phrase. La syntaxe regroupe les principes et les règles de construction des phrases dans une langue naturelle.

- la **sémantique** : étudie les signifiés c'est-à-dire la signification ou contenu du message.

- la **pragmatique** : étudie les rapports entre l'énoncé produit par des énonciateurs et la situation de cet énoncé. Ce sont les éléments du langage dont la signification ne peut être comprise qu'en connaissant le contexte.

Évidement, ces domaines et les règles associées ne sont pas cloisonnés. Il existe de nombreuses relations entre toutes ces approches et les différentes linguistiques. Ainsi, l'analyse lexicale (étude du lexique) ne peut être complètement conduite sans s'appuyer sur le sens des mots (sémantique). Cette analyse sémantique s'appuie elle-même sur la place et la fonction des mots dans la phrase (syntaxe) mais peut aussi s'appuyer sur le contexte de l'énonciation (pragmatique).

L'analyse automatique d'un texte est une manière d'obtenir une représentation de l'information contenue dans le texte par une compréhension informatisée de cette information. A cette fin, la rencontre de la statistique et de la linguistique donne naissance dès le début du XX$^{e}$ siècle à l'étude des textes via les nombres. Cette "nouvelle" approche en linguistique consiste à voir le document à travers la lorgnette des nombres et des chiffres. L'étude manuelle des distributions lexicales c'est-à-dire l'extraction de listes de mots (morphèmes et lexèmes) a permis à [Estoup, 1916] et plus tard à [Zipf, 1949] d'établir des lois empiriques. Georges Kinglsey Zipf a observé qu'en classant et en ordonnant les mots contenus dans un texte par ordre décroissant des fréquences d'apparition, on observe que le produit de cette fréquence par le rang du mot dans ce classement est égale à une constante et vaut approximativement 1 pour les textes longs (voir une démonstration accessible à l'adresse [1]).

$$\text{Rang x Fréquence} \simeq \text{constante K}$$

La notion de *"savoir ce que l'on compte"* est une notion importante qui apparaît dès lors que l'on désire compter et faire des analyses statistiques sur des objets que l'on veut comparer. Lebart appelle cela *"les unités de la statistique textuelle"* et précise qu'une norme doit être définie permettant d'isoler de la chaîne textuelle les différentes unités sur lesquelles porteront les dénombrements à venir. L'opération qui permet de découper le texte en unités minimales (c'est-à-dire en unités que l'on ne décomposera plus) s'appelle la segmentation du texte [Lebart et Salem, 1994].

Le choix de cette unité de comptage peut se faire à différents niveaux avec des complexités diverses :

- à l'échelle du mot pour obtenir le lemme ou le phonème.

- à l'échelle du syntagme pour obtenir la catégorie grammaticale.

1. http://users.info.unicaen.fr/~giguet/java/zipf.html

– à l'échelle de la phrase ou du texte pour isoler le(s) concept(s) et obtenir un sens.

Généralement le mot est pris comme unité de comptage dont la segmentation automatique consiste à extraire une suite de caractères délimités par des caractères délimiteurs (classiquement les signes de ponctuation, espace, etc.) à ses deux extrémités. Cependant, pour un même texte cette segmentation varie d'un domaine d'étude à l'autre et ne conduit pas au même décompte. Par exemple, le chercheur en informatique préférera le regroupement en une seule unité (le lemme) du singulier et du pluriel du mot "algorithme". Par contre ce regroupement ne sera pas souhaité dans l'étude des textes politiques. En effet, le pluriel d'un même substantif renvoie souvent à des notions différentes, parfois en opposition (par exemple l'opposition dans les textes récents de *défense de la liberté* / *défense des libertés* qui renvoie à des courants politiques opposés [Lebart et Salem, 1994]). Cette étape dite de **lemmatisation**, consiste à réduire chacun des mots en une entité appelée lemme (ou forme canonique). Ainsi la forme canonique d'un verbe est ce même verbe à l'infinitif, pour les autres mots la forme canonique est le mot au masculin singulier. Par exemple l'adjectif *"petit"* existe sous quatre formes *"petit", "petite", "petits" et "petites"*. Il existe beaucoup plus de formes du verbe "avoir" *(ai, as, a, avons, ais, avons eu, ayez eu, eussions eu, aurions, etc)*. Cette étape permet des décomptes sur des unités bien définies du point de vue de la "langue" et paraît séduisante. Cependant la lemmatisation d'un texte n'est pas une étape triviale et se heurte à des problèmes difficilement solvables. Il est souvent nécessaire de lever préalablement certaines ambiguïtés qui prennent naissance par exemple dans des homographies (*avions* issu du verbe *avoir*, et *avions* : substantif masculin pluriel), mais également dans des ambiguïtés sémantiques (Une livre de pain qu'il livre avec un livre de recette). Certaines de ces ambiguïtés peuvent toutefois être levées par une analyse grammaticale de la phrase ou l'examen du contexte immédiat ou éloigné (paragraphe, texte entier, etc.). L'intérêt de cette étape est largement discuté et Labbé écrit dans [Labbé, 2002] « l'expérience prouve que la correction orthographique, la normalisation des graphies et la lemmatisation sont des opérations indispensables pour mettre à disposition une information fiable ». Il évoque une expérience présentée en 1995 par M. Sylberztein sur deux années du journal Le Monde, expérience que résume la figure 4.1. Sur les articles de 1992, il a décompté 21,8 millions de mots sous près de 242 000 formes différentes. Puis l'année suivante il a décompté 23,2 millions de mots sous 247 000 formes. La comparaison du vocabulaire sur les deux années fait apparaître un noyau commun ridiculement faible. « Le journal aurait-il changé de langue entre 1992 et 1993 ? » Après correction et lemmatisation des graphies, le tableau est radicalement différent. Le vocabulaire ne compte plus "que" 41.000 entrées dont les deux tiers sont communes aux deux années. Cependant une vieille querelle a longtemps opposé les partisans et les adversaires de la lemmatisation [Brunet, 2000] et cette querelle semble trouver un consensus : « La décision, conclut André Salem, est d'ordre économique. Il est dans l'absolu toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes). Une lemmatisation complète, sur un corpus important, reste une opération coûteuse. Indispensable dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations et des typologies [...] ».

L'analyse statistique lexicale commence réellement son développement à partir de 1960 grâce à l'essor de l'informatique. Ainsi, le projet du *Trésor de la Langue Française* voit le jour et a pour vocation de constituer une grande bibliothèque informatisée, la base Frantext. Aidé par la constitution d'un tel corpus, Charles Muller entreprend une série d'études comparatives sur le vocabulaire des "grands auteurs" et étudie les œuvres de Corneille sur support infor-

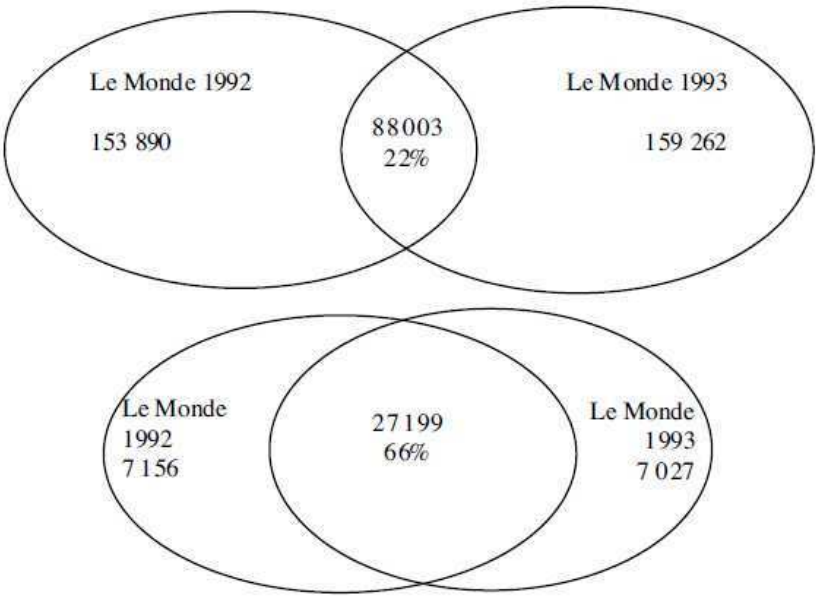



FIGURE 4.1 – Le vocabulaire du Monde sur deux années (en formes graphiques brutes puis en lemmes) ISSU DE [LABBÉ, 2002]

matique [Muller, 1967]. Ses méthodes quantitatives tentent de répondre à des préoccupations déjà anciennes sur la comparaison, la richesse, la spécificité ou l'évolution du vocabulaire entre différents auteurs. De nombreux travaux voient le jour où l'analyse stylistique permet de rechercher la paternité d'œuvres littéraires. Par exemple l'affaire Corneille-Molière [2] a fait et fait encore couler beaucoup d'encre. En effet, cette polémique de la littérature française commence avec le poète Louÿs qui attribue en 1919 l'œuvre de Molière à Corneille. Cette controverse sera reprise par des romanciers mais prendra une toute autre dimension lorsque Labbé la confrontera en 2001 aux statistiques textuelles [Labbé et Labbé, 2001]. Sa méthode, inspirée des travaux du domaine et de la distance intertextuelle introduite par E.Brunet [Brunet *et al.*, 2004], consiste à mesurer la distance entre deux textes. Dans [Brunet, 1988] la distance absolue entre deux textes **A** et **B** est la surface (en terme de vocabulaire) des deux textes moins leur intersection,

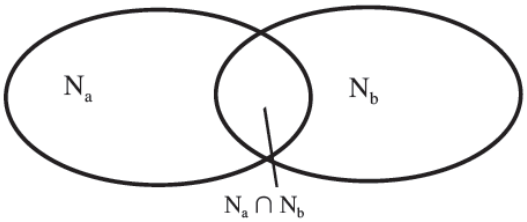



FIGURE 4.2 – Distance intertextuelle

2. http://corneille-moliere.org

$$\delta(a,b) = \frac{1}{2}\left(\frac{\sum_{i=1}^{N_a}|F_{ia} - F_{ib}|}{N_a} + \frac{\sum_{i=1}^{N_b}|F_{ib} - F_{ia}|}{N_b}\right) \tag{4.1}$$

c'est-à-dire la somme des différences entre les fréquences absolues de chacun des mots des deux textes (équation 4.1 et figure 4.2). Avec $F_{ia}$ et $F_{ib}$ fréquence du mot i dans A et dans B. $N_a$ et $N_b$ nombre de mots dans A et B (taille du texte). L'indice vaut 1 si les deux textes comparés ne partagent aucun mot. En revanche, le minimum théorique de cet indice ne peut atteindre zéro que dans le cas particulier de $N_a = N_b$. Ainsi dans [Labbé et Labbé, 2001], l'auteur propose de simuler la réduction du plus grand des deux textes à la taille du plus petit. Cette distance est appliquée sur les œuvres de Molière et de Corneille (64 pièces). L'auteur conclut qu'il y a une parenté d'une partie des pièces de Molière avec Corneille et qu'il s'opère un chassé-croisé entre les deux œuvres. Cette étude à été fortement critiquée par les statisticiens et les historiens.

La statistique lexicale s'est également bien développée dans les pays anglo-saxons où elle a surtout été utilisée en analyse stylistique. L'étude entreprise par Mosteller et Wallace [Mosteller et Wallace, 1984], dont les conclusions sur la paternité des textes de réflexion sur la nouvelle constitution américaine, **Le Fédéraliste "The Federalist Papers"**, furent convaincantes et reconnues par les disciplines historiques, donna une légitimité à ce domaine d'étude. **"The Federalist Papers"** est un recueil d'articles faisant la promotion de la nouvelle constitution américaine (1787-1788 ) et du nouveau gouvernement américain. Ces articles, écrits par James Madison, Alexander Hamilton et John Jay, signés sous le pseudonyme *Publius*, étaient une excellente référence pour comprendre la nouvelle Constitution américaine que le peuple était appelé à ratifier. Bien que la paternité de ces textes soit sûre pour 73 de ces articles, il reste un doute et une discorde entre les différentes écoles de pensée, sur la paternité des 12 autres textes. C'est l'analyse de ces textes par la statistique textuelle qui a permis de trancher sur l'attribution des auteurs [Mosteller et Wallace, 1984, Fung, 2003].

L'analyse lexicométrique a été utilisée sur les discours politiques [Lebart et Salem, 1994]. Dans [Labbé et Monière, 2000] les auteurs s'intéressent aux discours prononcés par les Premiers ministres québécois pour ouvrir les sessions parlementaires depuis 1945. L'utilisation de la distance inter-textuelle puis l'utilisation de la classification automatique (CAH) met en valeur quelques grands épisodes dans la vie politique de la province canadienne et souligne la singularité des deux passages au pouvoir du parti québécois. L'analyse lexicométrique des discours politiques se retrouve dans de nombreux travaux [Marchand et Monnoyer-Smith, 2000], [Labbé et Monière, 2008], [Foucault et Francois, 2009].

Jean-Paul Benzécri [Benzécri et Benzécri, 1980] ouvre en France, le champ de l'analyse et l'exploration des données multidimensionnelles ainsi que des méthodes de classification automatique [Sebastiani, 2002]. Ces travaux ont permis l'analyse de grands tableaux de contingence grâce à l'analyse des correspondances et à l'analyse factorielle. Ces analyses permettent de tester l'indépendance des données et de décrire les proximités ou éloignement des données contingentées (sous l'hypothèse d'indépendance). Ainsi, il est possible de tracer la cartographie des associations lexicales susceptibles de révéler les réseaux sémantiques[3] ou modèles

3. représentation des concepts (portant un sens dans un domaine donné) au travers de leurs relations mutuelles.

mentaux[4] dont le texte porte la trace. C'est ensuite au spécialiste d'interpréter ces entités abstraites. Cela permet par exemple de retrouver les grandes thématiques dans un texte et de mettre à jour les structures linguistiques qui les caractérisent. Dans [Marchand, 2008] une carte d'Analyse Factorielle des Correspondances (AFC) (voir figure 4.3) est construite à partir de l'analyse des déclarations de politique générale des Premiers ministres sous la $V^{ieme}$ république. Cette représentation permet d'identifier l'évolution des déclarations politiques en fonction du temps et permet de voir la répartition des discours politiques en fonction des proximités et des éloignements de leur lexique. *Après la dernière élection présidentielle (6 mai 2007), le nouveau président a convoqué les parlementaires de sa majorité avant la déclaration de politique générale de son premier ministre, pour leur livrer sa « feuille de route ». Les commentateurs ont été unanimes pour dire qu'il avait prononcé un discours de politique générale. Or, la figure 4.3 montre que ce discours ne partage pas les propriétés d'une déclaration de politique générale : les procédures statistiques (AFC) isolent ce discours et l'opposent à l'ensemble des autres déclarations de la $V^{ieme}$ République* [Marchand, 2008].

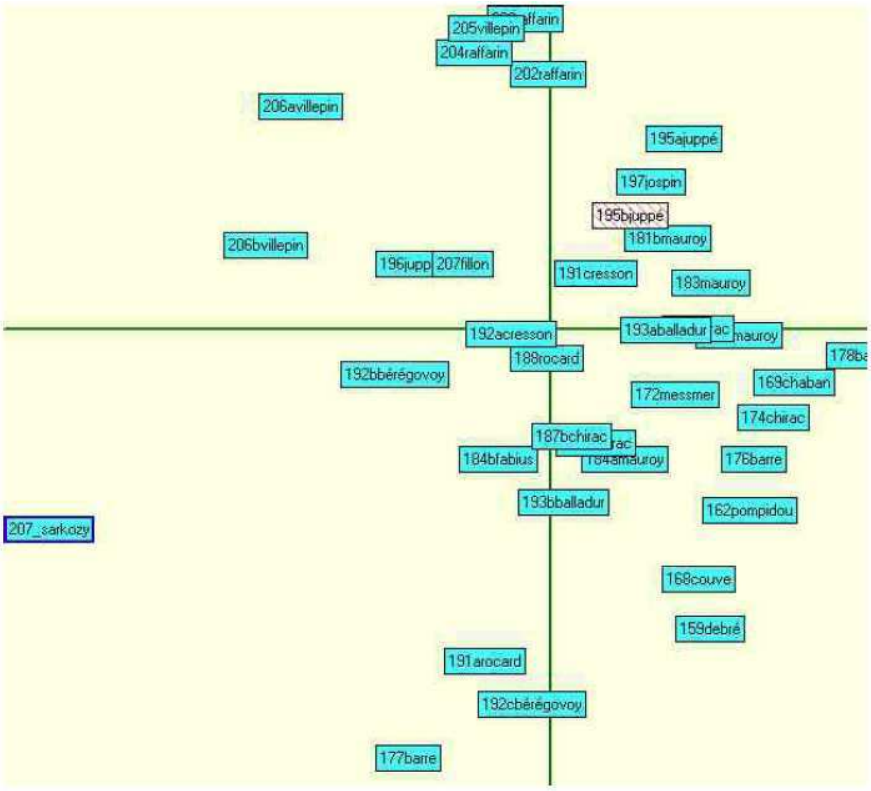



FIGURE 4.3 – AFC des déclarations de politique générale des Premiers Ministres français (1959-2007) et de la feuille de route de N. Sarkozy. ISSU DE [MARCHAND, 2008]

Cette analyse exploratoire multidimensionnelle couplée à l'analyse textuelle est largement utilisée dans les études qualitatives comme le dépouillement de questionnaires dans des études en sociologie [Patrice *et al.*, 2004, Guérin-Pac, 1997], psychologie et marketing [Desmarais et Moscarola, 2004]. Dès l'apparition du courant *"Benzécriste"* de nombreuses approches informatisées pour l'analyse statistique des textes ont vu le jour comme SPADT (Le-

4. modèles qui combinent des représentations du monde, des modes de raisonnement, des guides pour l'action.

bart, ENST et CISIA, Paris), LEXICO (Salem, Paris 3), HYPERBASE (Brunet, Université de Nice) et ALCESTE (Reinert, CNRS). La méthode ALCESTE, à l'origine du logiciel du même nom [Reinert, 1986], permet de quantifier un texte pour en extraire les structures signifiantes les plus fortes. Selon son auteur [Reinert, 1997, Reinert, 2002, Reinert, 2008] cette méthode permet à travers un ensemble de calculs, de cartographier les principaux lieux communs d'un discours, ou mondes lexicaux, qui sont des traces purement sémiotiques inscrites dans la matérialité même du texte. Un "monde lexical" est la trace lexicale d'un "référent" ou "point de vue" particulier utilisé par l'énonciateur pour construire ses énoncés. ALCESTE procède par fractionnements successifs du texte en fragments de tailles relativement analogues, nommés "unités de contexte" (séquences de textes de longueurs comparables, qui peuvent souvent coïncider avec les phrases). Ces fragments sont ensuite classés statistiquement selon une procédure descendante hiérarchique. L'objectif de cette classification descendante hiérarchique est la répartition des énoncés en classes marquées par le contraste de leur vocabulaire [Kalampalikis, 2003]. Elle a pour avantage de ne pas exiger de connaissances *a priori* sur le texte à analyser.

#### 4.1.1.1 L'évolution vers la topologie textuelle

L'Analyse de Données Textuelles (ADT), on l'a vu, est basée sur les dénombrements et les fréquences des mots. Cette analyse voit le texte comme un sac de mots et renonce à prendre en compte le positionnement dans le texte des unités dénombrées. Aujourd'hui cet état de fait évolue et Mayaffre dans [Mayaffre, 2007] constate que : « l'analyse de données textuelles admet aujourd'hui qu'un texte ou un corpus textuel n'est pas seulement une urne anarchique pleine de données linguistiques mélangées, mais aussi un espace ou un plan sur lequel ces données s'enchaînent (plus que s'additionnent) et s'organisent au fil du texte ». C'est Etienne Brunet qui le concède en 2006 après une vie de recherche consacrée à l'ADT. Malgré les travaux pionniers de Lafon en 1984 [Lafon et Muller, 1984] sur les rafales, il regrette en effet que : « l'ADT se soit surtout attachée jusqu'ici aux fréquences, sans trop s'occuper des séquences ». Désormais l'organisation non-séquentielle originelle de la lexicométrie prend en compte l'organisation spatiale, linéaire ou continue et devient l'organisation topologique textuelle.

C'est Lamalle et Salem qui les premiers font une "cartographie textuelle" en localisant au fil du corpus l'apparition des unités linguistiques (les mots) retenues [Lamalle et Salem, 2002]. Dans cette étude d'où est extraite la figure 4.4, les 2319 paragraphes du corpus des congrès de la CFDT entre 1978 et 1998, sont représentés par un carré de taille fixe. Les carrés de couleur sombre signalent la présence, au sein du paragraphe concerné, d'au moins une occurrence du type cartographié (ici la classe **nego**+ contient un ensemble de mots en rapport avec la négociation). Ce type de représentation permet de localiser et de voir les périodes concernées par cette unité de décompte.

La linéarité du texte et ses enchaînements thématiques ne sont plus ignorés dans ces analyses. Une représentation comme celle de la figure 4.4 a un pouvoir descriptif évident qui permet la compréhension des corpus textuels en restituant leur déroulement. Dans [Salem, 2004], l'auteur fait apparaître grâce au lexique les liens entres deux textes analysés simultanément. Lorsque les deux textes sont alignés (cet alignement peut se faire de plusieurs façons : en se basant sur les paragraphes ou à partir d'une chronologie ou bien encore dans le cas de discours en fonction des tours de paroles, etc.) il est possible, en dressant la topologie des deux textes,

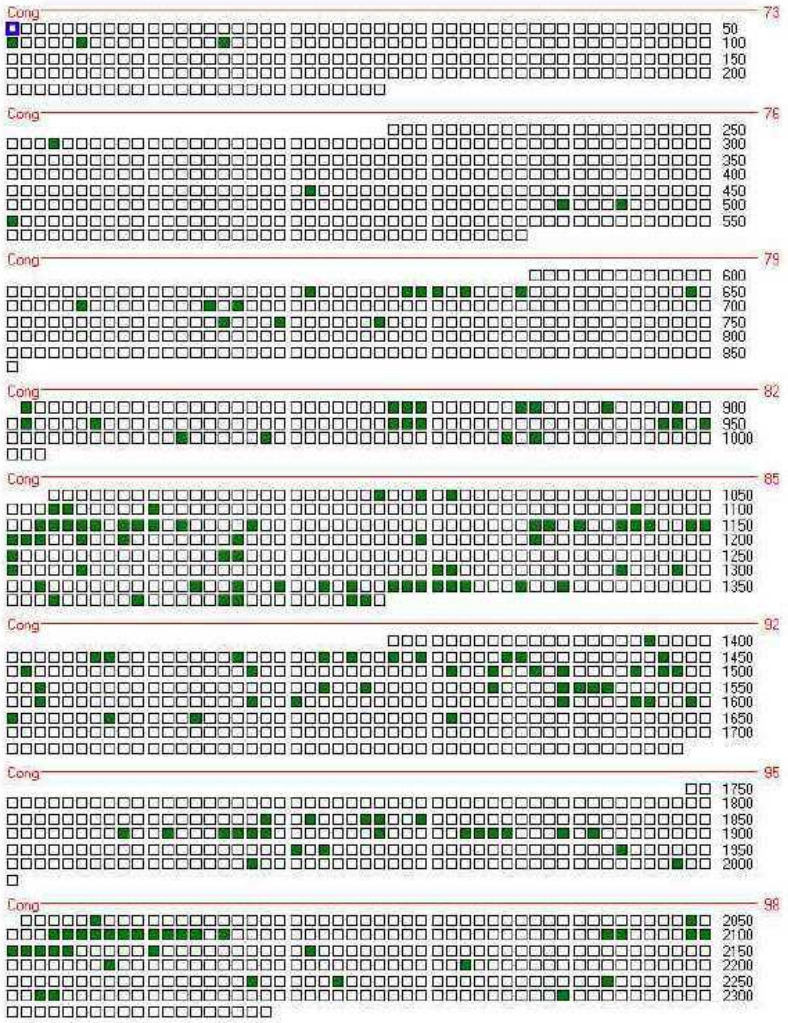



FIGURE 4.4 – Topologie des occurences de la classe **nego+** dans les 2319 paragraphes des 8 congrès CFDT (*La division du texte en périodes (années des congrès de l'organisation syndicale) est matérialisée par des lignes horizontales rouges.* ISSU DE [LAMALLE ET SALEM, 2002]

de suivre l'évolution des vocables [5] choisis et de suivre les interactions entre les deux textes ; ce phénomène est nommé par l'auteur *la résonance textuelle*. L'auteur présente comme exemple une confrontation verbale entre plusieurs locuteurs (F. Mitterrand et J. Chirac en 1988), la résonance permet de juger de l'influence des productions de chacun des locuteurs sur celles de l'autre. Dans [Longrée *et al.*, 2004], les auteurs étudient la répartition globale des temps verbaux le long de l'axe syntagmatique [6] pour évaluer les distances et similarités entre des textes lemmatisés d'historiens latins.

L'utilisation du voisinage d'un vocable pris comme pivot permet l'analyse des micro-

5. Élément du langage, considéré quant à sa signification et à son individualité lexicale, synonyme de mot. (DÉFINITION DU TLFI)

6. axe horizontale sur lequel s'opère l'enchaînement de l'énoncé (la chaîne parlée / écrite)

distributions. Dans [Viprey, 2004], cette micro-distribution est repérée dans le temps et redistribuée sous contrainte d'équidistribution. L'utilisation d'un analyse par AFC permet ensuite d'obtenir une vue synthétique de l'évolution du vocable pivot et de ses cooccurrents. Certains auteurs s'intéressent à la mesure automatique de ces caractéristiques spatiales. Dans [Brunet, 2006] l'auteur s'intéresse aux séquences dont la représentation graphique permet de repérer les rafales et le rythme du discours. Son approche consiste à suivre le parcours (en terme de fréquence) d'un mot tout au long du corpus (constitué d'une compilation de textes littéraires) sans s'arrêter aux barrières des textes. La figure 4.5 illustre ceci. L'axe horizontal représente le corpus où les textes et leurs frontières sont représentés par des lignes verticales. Le mot recherché est représenté séquentiellement de sa première occurrence (en bas à gauche) à sa dernière (en haut à droite), chaque point étant déterminé en abscisse par la position du mot dans le corpus et en ordonnée par le numéro de l'occurrence. Quand les points se concentrent et s'orientent vers la verticale, il s'agit d'une *"rafale"*, une concentration des occurrences due à un changement thématique ou stylistique. Quand les points s'espacent et s'inclinent à l'horizontale, cela correspond à une raréfaction momentanée de l'objet recherché. Dans le corpus romanesque traité par Brunet, l'amour est utilisé dans la figure 4.5 pour illustrer la méthode. L'amour n'est pas équitablement partagé. « Il jaillit verticalement dans la Nouvelle Héloïse, dans Indiana de Georges Sand et dans Un Amour de Swann et il s'étiole en un mince filet languissant de Flaubert à Verne et Zola ». L'auteur utilise ensuite une métrique (le test de *Lafon*) lui permettant de retrouver les distributions irrégulières. Cette analyse est réitérée sur un ensemble de mots permettant ainsi de lister les vocables dont les répartitions sont régulières ou au contraire irrégulières dans l'espace ou le temps.

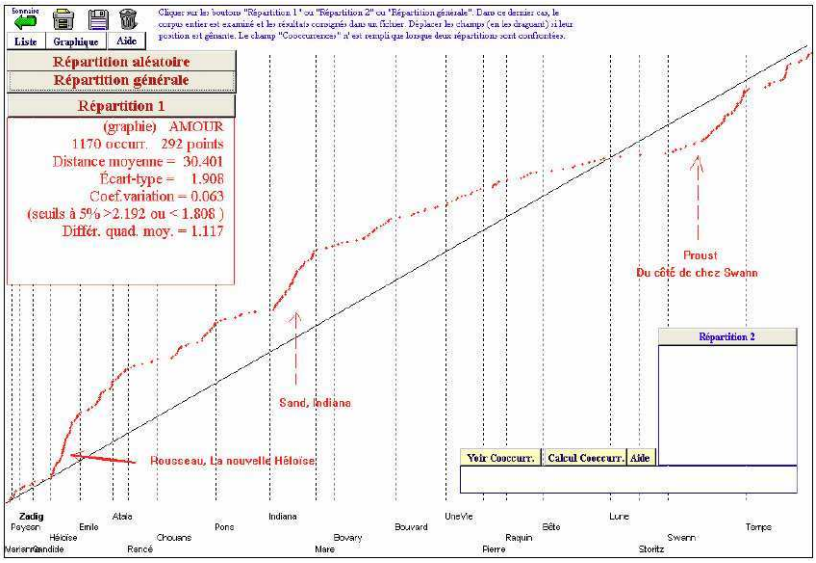



FIGURE 4.5 – Représentation graphique en séquence du mot amour, de Marivaux à Proust ISSU DE [BRUNET, 2006]

Pour [Vonfelt, 2008] l'axe diachronique du texte est vu comme un axe temporel sur lequel apparaissent les occurrences de l'unité linguistique comptée (dans cette étude ce sont les caractères). La mesure du "temps" séparant l'apparition de ces occurrences prises deux à deux (appelée *"temps de retour"*) sert à caractériser "l'activité" de l'unité linguistique. Pour mesurer la distance entre deux textes Vonfelt calcule la distribution puis la répartition de ces temps de retour pour plusieurs unités linguistiques. La distance entre deux textes est obtenue

en comparant les courbes de répartition pour un ensemble d'unités linguistiques.

L'analyse de données textuelles permet d'extraire, de représenter et de synthétiser l'information contenue dans de gros corpus textuels. La puissance de l'outil informatique couplé à la statistique permet d'appréhender plus facilement et en un minimum de temps cette masse de données que représente le texte. Dans cette thèse, nous allons nous intéresser aux informations textuelles contenues dans les fiches d'inscription et plus particulièrement aux synopsis des films présentés au FIFA que nous avons déjà présentés dans le chapitre 2.1.3.

## 4.2 La statistique textuelle appliquée aux synopsis des films d'animation

Pour inscrire un film à la sélection du festival, les auteurs doivent fournir une fiche d'inscription contenant des informations concernant leur œuvre. Nous allons travailler sur les synopsis de ces films qui sont des textes courts (en moyenne 20 mots, voir figure 4.6) en français et en anglais et qui décrivent sous forme de résumé ou d'accroche le sujet traité par le film. Notre corpus est constitué des synopsis des films d'animation inscrits au festival et des films d'animation hors concours issus d'un fond historique (films du début du XX$^{e}$ siècle). La figure 4.6 présente un bilan lexical des différents champs des fiches d'inscription.

| | Non-réponses | Nombre de mots | Nombre moyen de mots | Nombre de mots différents | Nombre de mots uniques | Fréquence maximum | Mot le plus fréquent |
|---|---|---|---|---|---|---|---|
| Titre français | 12750 | 17444 | 3.23 | 6687 | 4982 | 41 | histoire |
| Titre anglais | 16092 | 6550 | 3.17 | 2401 | 1824 | 27 | Life |
| Synopsis (français) | 0 | 420179 | 23.14 | 31060 | 15700 | 1445 | homme |
| Synopsis (english) | 4438 | 309418 | 22.56 | 23418 | 11982 | 1141 | world |
| Synopsis (français) lémmatisé | 1 | 221514 | 12.20 | 24018 | 11985 | 4478 | être |

FIGURE 4.6 – Statistiques sur le corpus textuel issu des 18155 fiches d'inscription du CITIA

On peut voir un certain nombre de caractéristiques de ces textes. On remarque dans cette base que les titres des films ne sont pas souvent renseignés. Ceci s'explique par la nature de cette base qui contient les fiches de films hors concours (non formatées pour le FIFA). C'est par exemple le cas de films publicitaires qui n'ont pas de nom. Les synopsis anglais ne sont pas toujours disponibles (absents pour 24% de la base). Ceci s'explique par le fait que ces textes sont souvent issus d'une traduction des synopsis français. Le lexique utilisé dans les titres des films est un vocabulaire riche puisque les mots différents représentent 40% de l'ensemble des mots utilisés dans les titres. Ce chiffre descend à 8% dans le cas des synopsis. Ceci traduit le fait que les synopsis sont des textes descriptifs utilisant peu de vocabulaire spécifique. Ce vocabulaire est consensuel puisque partagé par l'ensemble des synopsis. On retrouve cette tendance lorsque l'on analyse les mots uniques (hapax[7]). En effet, 30% des mots utilisés dans les titres sont des hapax c'est-à-dire qu'ils ne sont pas partagés par d'autres titres. Seulement 4% des mots utilisés dans les synopsis sont spécifiques et utilisés une seule fois (zoolympique, compacteur, incorruptible, émérite, percepteur, Benayoun, etc.). Le calcul des mots les plus fréquents est réalisé à partir des textes débarrassés des mots outils (mots

7. hapax ou apax désigne un fait de langue (mot, expression, construction) dont il n'existe qu'une seule occurrence dans un corpus donné. LAROUSSE

vides de sens, voir la définition juste après §4.2.1.1). On voit que le terme le plus récurrent dans les synopsis français est **homme** et **world** dans les synopsis anglais. Cette différence peut s'expliquer par la différence de taille des deux corpus (le corpus anglais est 25% moins grand que celui du français) et surtout par la traduction du mot homme suivant que l'on parle de la personne ou de l'humanité (man, mankind, human, humanity, etc.).

Dans la suite de cette étude, nous allons nous intéresser au corpus textuel constitué des 18155 synopsis français. Ce corpus contient un peu plus de 420 000 mots ce qui fait que les analyses statistiques ont du *sens*. En effet, plus le corpus est volumineux, plus les régularités et effets statistiques sont significatifs.

### 4.2.1 Analyse lexicale globale

Puisque la masse de données textuelles est volumineuse, une analyse et une approximation lexicale s'impose afin d'en extraire les informations significatives. L'idée est d'appréhender le texte à partir des mots et/ou expressions les plus fréquemment utilisés. Dans un premier temps nous analysons la macro-distribution des termes dans le corpus. Cette analyse lexicale est réalisée en utilisant le logiciel "Le Sphinx" [Sphinx, 2009].

#### 4.2.1.1 Réduction du lexique

**4.2.1.1.1 Mots vides** Afin de réduire la taille du lexique et rendre ainsi son interprétation plus facile il est nécessaire de supprimer un certain nombre de mots n'apportant pas réellement de sens. Ces **mots outils** ou *mots grammaticaux*, sont des mots dont le rôle syntaxique et grammatical est plus important que le rôle sémantique (je, et, en, le, etc.). Mais plus généralement ces **mots vides** (ou "stop words", en anglais) sont des mots qui sont tellement communs qu'il est inutile de les conserver car non discriminants pour le document (de, le, sa, maintenant, encore, etc). La détection puis la suppression de tels mots à partir d'un dictionnaire de mots vides permet de réduire la taille du lexique (voir le tableau 4.1 dans la liste du bas où les mots "de" et "le" etc. ont été supprimés). Nous remarquons également dans le texte réduit (sans mots outils) l'apparition du terme *"dun"* qui est en réalité une omission de l'apostrophe dans le fichier source.

Après suppression des mots-outils nous obtenons une liste de mots caractérisant macroscopiquement le corpus. On remarque que les termes comme *homme*, *enfants*, *femme*, *fille*, *garçon* renvoient aux personnages mis en scène dans les films. Ces personnages sont des êtres humains. Les termes comme *vie*, *monde*, *histoire*, *série*, *aventures*, *amour*, renvoient au contexte et à l'histoire du film. Puis on a un ensemble de qualificatifs comme *petit*, *jeune*, *petite*, *grand* qui permettent d'apporter un complément d'information. On remarque la présence fréquente de termes pluriels comme *enfants* et *aventures* signifiant que les synopsis traitent le plus souvent de plusieurs enfants et de plusieurs aventures. On remarque également que *petit* est deux fois plus utilisé que *petite*, ce constat est aussi vrai pour *grand* et *grande*. Les synopsis emploient-ils plus souvent le masculin que le féminin ?

**4.2.1.1.2 Segments répétés** De tels aperçus de textes, portés par les lexiques, peuvent parfois conduire à de mauvaises interprétations. En effet, certaines associations de mots sont nécessaires pour comprendre quel est le signifié et résoudre des ambiguïtés (par exemple

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Texte Brut** | de | 20477 | un | 13289 | la | 12535 | | |
| | et | 11377 | le | 9456 | une | 8958 | | |
| | a | 8910 | les | 7296 | l | 7010 | | |
| | d | 6886 | des | 5910 | dans | 5275 | | |
| | en | 4617 | est | 4276 | pour | 3995 | | |
| | du | 3934 | il | 3830 | qui | 3697 | | |
| | se | 3514 | sur | 3084 | son | 3018 | | |
| | par | 2584 | au | 2273 | sa | 2055 | | |
| | s | 1821 | que | 1809 | avec | 1693 | | |
| | ce | 1581 | mais | 1528 | ses | 1509 | | |
| **Texte sans les mots-outils** | homme | 1445 | film | 1389 | vie | 1387 | | |
| | monde | 1268 | histoire | 1189 | deux | 1119 | | |
| | petit | 1084 | série | 768 | fait | 732 | | |
| | jeune | 692 | enfants | 687 | femme | 601 | | |
| | aventures | 588 | après | 583 | ville | 575 | | |
| | petite | 570 | être | 557 | fille | 540 | | |
| | jour | 524 | très | 523 | faire | 507 | | |
| | garçon | 491 | même | 478 | peut | 471 | | |
| | grand | 469 | trois | 466 | temps | 465 | | |
| | publicité | 453 | dun | 447 | amour | 440 | | |

TABLE 4.1 – Liste des 30 mots les plus fréquents dans le corpus brut et réduit (sans les mots-outils) avec les nombres d'occurrences pour chaque mot

le mot "arrêter" seul est ambigu alors que "arrêter de travailler" ne l'est pas). Ainsi, il faut restituer chaque mot dans son contexte en cherchant les segments répétés et en produisant des cartes d'associations lexicales. Les segments répétés (séquences de mots répétés à l'identique) renvoient les rigidités du texte, comme les formules toutes faites ou les expressions.

La figure 4.7 montre la liste des 100 premiers segments répétés. On retrouve des expressions de la langue française comme : *mettant en scène*, *peut-être*, *était une fois*, *jusqu'au jour*, mais également des éléments du genre et de la technique du film d'animation comme : *Spot publicitaire*, *dessin animé*, *Bande-annonce*, *science-fiction*, *noir et blanc*, *pâte à modeler*, *vues réelles*, etc. Ces éléments techniques permettent de décrire le contexte ou la technique du film. Voici quelques synopsis dans lesquels apparaissent les termes précédents :

- *Film publicitaire pour l'eau de Vittel.*
- *Film publicitaire pour les lampes Mazda.*
- *Film publicitaire pour Lucky Strike.*
- *Film publicitaire pour les bonbons Mentos.*
- *Court-métrage pilote présentant deux super héros au chômage.*
- *Ce court-métrage d'animation raconte la cavale du criminel à travers la ville.*
- *Au cours d'une projection d'un court métrage européen, le film devient un vrai casse-tête pour le projectionniste.*
- *Une main dégrossit, défriche la matière blanche de la pâte à modeler. Elle joue, elle caresse au gré de ses caprices, pour la simple beauté du geste et, petit à petit, à force d'empreintes, façonne un visage.*

| | | |
|---|---|---|
| petit garçon (194) | jusqu'au jour (46) | Porky Pig (30) |
| petite fille (185) | Séquence animée (46) | prend vie (30) |
| mettant en scène (166) | dessins animés (45) | monde imaginaire (29) |
| jeune fille (133) | père Noël (45) | premiere fois (29) |
| Spot publicitaire (114) | met en scène (45) | Looney Tunes mettant en scène Porky (29) |
| dessin animé (98) | petite ville (41) | jeunes enfants (28) |
| Looney Tunes (98) | pâte à modeler (41) | bande dessinée (28) |
| Spot publicitaire (98) | personnage principal (41) | prises de vues (28) |
| peut-être (96) | raconte l'histoire (41) | part à la recherche (28) |
| Looney Tunes mettant (96) | vues réelles (39) | était une fois (28) |
| Looney Tunes mettant en scène (95) | tombe amoureux (39) | scène Porky Pig (28) |
| jeune garçon (91) | Looney Tunes mettant en scène Bosko (38) | chaperon Rouge (27) |
| jeune homme (90) | prennent vie (37) | mettant en scène Porky Pig (27) |
| Merrie Melodies (86) | film montre (36) | voyage à travers (27) |
| vieil homme (80) | grande ville (36) | après-midi (26) |
| Film publicitaire (78) | homme et une femme (36) | Bande-annonce (26) |
| grand-mère (74) | Séquence animée (36) | êtres humains (26) |
| Bande-annonce (66) | vieille dame (36) | Histoire dun (26) |
| Film publicitaire (66) | deux amis (34) | histoire d'un homme (26) |
| Histoire d'amour (65) | deux hommes (34) | Trois Petits (26) |
| grand-père (64) | épisode de la série (34) | Looney Tunes mettant en scène Porky Pig (26) |
| jeune femme (63) | petit village (33) | monde entier (25) |
| aujourd'hui (62) | petit bonhomme (33) | deux jeunes (24) |
| film d'animation (61) | Séquence animée pour MTV (33) | Jeu vidéo (24) |
| était une fois (58) | moyen âge (32) | vieille femme (24) |
| deux enfants (56) | film noir (32) | courts métrages (23) |
| vie quotidienne (56) | petit homme (32) | petit déjeuner (23) |
| deux personnages (55) | conte de fées (31) | faire face (23) |
| jusqu'au jour (51) | dun homme (31) | guerre mondiale (23) |
| Non communiqué (51) | scène Porky (31) | Petit Chaperon (23) |

FIGURE 4.7 – Liste des 100 premiers segments répétés rangés par nombre d'occurrences. *Les expressions commençant par une majuscule sont soit des noms propres, soit des expressions qui commencent au moins une fois dans le corpus le début d'une phrase.*

- *La réalité, le sexe et l'amour font peur à Ana qui manque d'expérience et tombe sous le charme d'une star de cinéma érotique en pâte à modeler.*
- *Peinture abstraite animée en pâte à modeler.*

On retrouve également dans ces segments répétés des personnages de contes et de célèbres maisons de production comme *chaperon Rouge*, *Porky Pig*, *Looney Tunes*, *Merrie Melodies* ceci vient probablement du fait qu'un certain nombre de films sont antérieurs au festival et correspondent aux films hors concours (films anciens créés avant le festival).

**4.2.1.1.3 Lemmatisation** Pour simplifier le lexique et augmenter la pertinence de son analyse il peut être intéressant de supprimer les formes fléchies de certains mots. Cette étape dite de **lemmatisation**, consiste à réduire chacun des mots en une entité appelée lemme (ou forme canonique). De ce fait, on rassemble par exemple les occurrences de *petit, petite, petits, petites* avec celle de *petit* ou bien encore les occurrences de *avoir, ai, as, a, avons, ais, avons eu, ayez eu, eussions eu, aurions, etc.* avec celle de *avoir*. Chacun des mots du lexique est remplacé par sa forme canonique (voir tableau 4.2 le mot "petite" est devenu "petit", le mot "enfants" est devenu "enfant").

Pour les substantifs, lorsque l'on regarde les 30 termes les plus fréquents dans le corpus lemmatisé et réduit on retrouve un lexique dont les caractéristiques sont les mêmes que celui du corpus sans la lemmatisation. On retrouve un vocabulaire lié à l'être humain (homme, femme, enfant, fille) plus un terme *personnage* qui apparaît. On retrouve les ingrédients d'une histoire avec une quête *aventure*, *découvrir*, *trouver*. Par contre on constate bien un phéno-

| | | | | | | |
|---|---|---|---|---|---|---|
| **Texte Lemmatisé sans mots-outils** | être | 4478 | petit | 1749 | homme | 1660 |
| | film | 1529 | histoire | 1358 | vie | 1306 |
| | faire | 1214 | monde | 1205 | avoir | 1173 |
| | pas | 1138 | enfant | 1059 | plus | 951 |
| | pouvoir | 910 | tout | 908 | jeune | 864 |
| | grand | 746 | aventure | 737 | femme | 706 |
| | personnage | 700 | voir | 678 | autre | 677 |
| | aller | 668 | jour | 625 | série | 612 |
| | fille | 586 | ami | 573 | ville | 570 |
| | découvrir | 541 | très | 523 | trouver | 522 |

TABLE 4.2 – Liste des 30 mots les plus fréquents dans le corpus lemmatisé et réduit avec les nombres d'occurrences pour chaque mot

mène bien connu de la lemmatisation qui est de faire remonter les verbes. Ainsi le verbe *être* devient le mot le plus fréquent traduisant une description de l'état de l'histoire (personnage, contexte, etc.). Le verbe *faire* est ensuite le verbe le plus fréquent, il traduit une description de l'action de l'histoire (les personnages *font* quelque chose). Le verbe *avoir* est le troisième verbe le plus fréquent, il traduit une description des propriétés (le personnage *a* quelque chose).

#### 4.2.1.2 Associations lexicales

D'une manière moins rigide que les segments répétés la statistique des associations lexicales (via l'AFC) donne une idée de la propension à associer les mots les uns aux autres ou au contraire à ne pas les faire coexister dans une même expression en l'occurrence le synopsis. Cette analyse est intéressante lorsque l'on ne sait pas quelles sont les grandes thématiques du corpus. Pour réaliser cette analyse, nous reprenons le lexique du tableau 4.2 des 100 lemmes les plus fréquents. Ainsi, chaque synopsis ou entrée du corpus est décrit par un sous ensemble de ces 100 termes. L'analyse des associations lexicale de ces termes consiste à créer un tableau de contingence (voir matrice 4.2), où $n_{i,j}$ représente le nombre de synopsis contenant à la fois le terme i et le terme j. Plus ce nombre est important plus les termes ont tendance à être associés dans un synopsis.

$$\begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,J} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ n_{I,1} & n_{I,2} & \cdots & n_{I,J} \end{bmatrix} \tag{4.2}$$

Une AFC permet ensuite de déterminer s'il existe un lien privilégié ou non entre 2 termes. Cette analyse compare les effectifs du tableau à ceux qu'on aurait obtenus si les effectifs étaient répartis proportionnellement et indépendamment. Pour tester cette hypothèse d'indépendance, le test du $Chi^2$ consiste à mesurer l'écart entre ce qui est constaté et le cas d'indépendance. Si la mesure du $Chi^2$ est grande on présume l'existence entre les deux modalités d'un lien d'autant plus significatif que l'écart est grand.

On peut donner une représentation plus visuelle des écarts à l'indépendance par l'utilisation d'une carte d'analyse factorielle des correspondances. Elle consiste à tracer une carte à

partir des résultats de l'AFC en disposant les modalités en fonction des écarts à la situation d'indépendance. Par défaut, chaque modalité est représentée par un pavé de surface proportionnelle à son effectif. Leurs positions les unes par rapport aux autres s'interprètent ainsi (issu du manuel du logiciel Sphinx) :

- Deux modalités lignes et colonnes seront d'autant plus proches que les effectifs du tableau sont en excès par rapport à l'indépendance : attraction.
- Les modalités lignes et colonnes seront d'autant plus éloignées que les effectifs du tableau sont en déficit par rapport à l'indépendance : répulsion.
- Les modalités lignes ou colonnes situées à la périphérie de la carte signalent des profils originaux. Au contraire, une position centrale interdit tout commentaire (profil sans originalité ou point mal représenté dans le système d'axes de la carte).

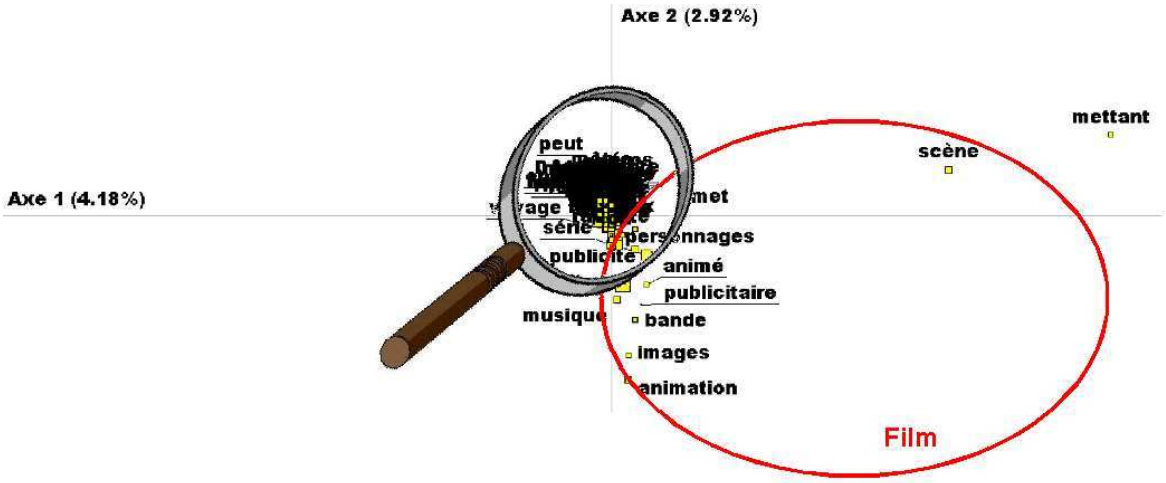



FIGURE 4.8 – Carte de l'Analyse factorielle des correspondances multiples pour la recherche de thématiques dans le corpus sans lemmatisation.

Les cartes AFC des figures 4.8 et 4.9 (zoom de la partie centrale de la figure 4.8) calculées sur les synopsis français sans mots outils permettent, à travers l'agencement des modalités et des constellations, d'identifier des réseaux sémantiques (ou configurations sémantiques). Nous pouvons identifier à partir de ces cartes 4 constellations sémantiques permettant d'identifier des thématiques propre au corpus :

- La thématique du **Film** contenant les mots clefs *"film, images, musique, bande, animation, animé, publicité, publicitaire"* (sur la figure 4.8) regroupe les éléments typiques de construction des films d'animation. Ces mots clefs sont assez éloignés du reste du lexique (origine de la carte) montrant le coté atypique de ces termes et de cette thématique.

- La thématique **Les Histoires** contenant les mots clefs *"série, aventures, raconte, nouvelle, conte, voyage, histoire, monde, personnage, personnages, héros, enfants"* (quadrant du bas de la figure 4.9). On retrouve les caractéristiques des types d'histoires racontées avec dans la partie de gauche une proximité entre *conte*, *raconte*, *nouvelle* qui correspond au type d'histoire et une proximité plus dans la partie de droite entre *personnage(s)*, *Héros*, *animaux*, *enfants*.

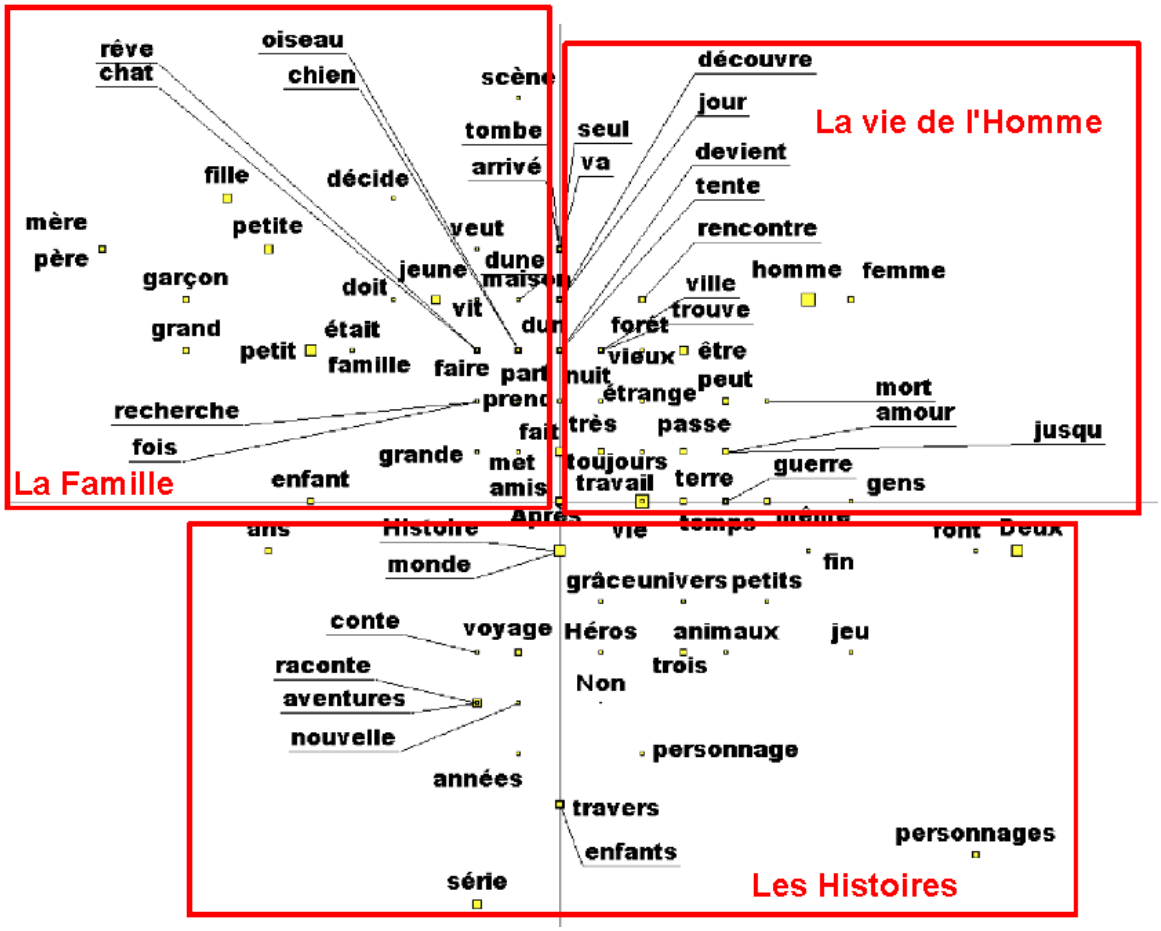



FIGURE 4.9 – Carte de l'Analyse factorielle des correspondances multiples pour la recherche de thématiques dans le corpus sans lemmatisation (zoom sur la constellation au centre du graphique).

- La thématique **La vie de l'Homme** contenant les mots clefs *"homme, femme, rencontre, vieux, amour, mort, travail, terre, guerre, gens"* (quadrant haut-droit de la figure 4.9). On remarque la proximité entre *amour* et *mort* et donc de leurs cooccurrences fréquentes, ce qui montre que l'on est dans un registre dramatique.

- La thématique **La Famille** contenant les mots clefs *"famille, mère, père, garçon, fille, enfant, maison, chien, chat, oiseau"* (quadrant haut-gauche de la figure 4.9). Ce groupe est associé à différents qualificatifs comme *petit(e)*, *grand(e)*, *jeune*.

Notons que la variance totale expliquée par les axes principaux retenus semble assez faible. Ceci est normal puisque nous travaillons sur des données textuelles associées aux synopsis (une vingtaine de mots en moyenne) dont l'unité de comptage est le mot. Le nombre de combinaisons possibles entre les termes retenus au sein d'un synopsis est donc quasi infini augmentant ainsi considérablement les dimensions de l'espace de représentation des associations. Aussi, la projection de cet hyperespace dans un espace bidimensionnel se traduit par une variance totale expliquée assez faible. De plus cette approche ne tient absolument

pas compte des relations spatiales qui existent entre les mots (relations essentiellement dues à la syntaxe). Finalement les synopsis sont vus comme un sac de mots et les thématiques identifiées ne sont que des tendances à associer des termes entre eux au sein des synopsis. De plus l'interprétation de ces regroupements n'est pas naïve et reste guidée par une certaine expertise du corpus et du domaine. Malgré tout il est intéressant de voir si l'on retrouve ces mêmes thématiques sur le corpus lemmatisé.

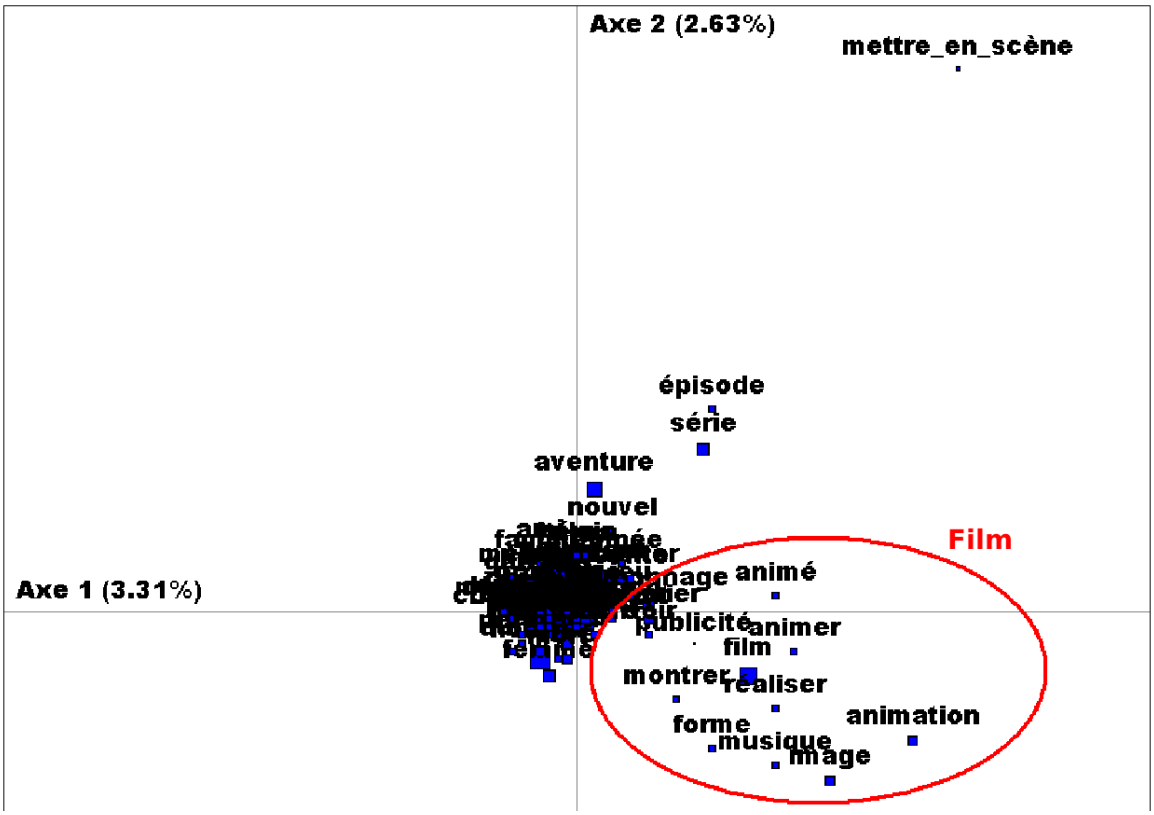



FIGURE 4.10 – Carte de l'Analyse factorielle des correspondances multiples pour la recherche de thématiques dans le corpus avec lemmatisation

Les cartes AFC des figures 4.10 et 4.11 (zoom de la partie centrale de la figure 4.10) calculées sur les synopsis français réduits aux lemmes et aux segments répétés sans les mots-outils, permettent à travers l'agencement des modalités et des constellations d'identifier les mêmes réseaux sémantiques que précédemment :

- La thématique du **Film** contenant les mêmes mots clefs que précédemment avec en plus *"montrer, réaliser, animer"* (sur la figure 4.10) regroupe les éléments de construction des films d'animation. L'ajout de ces mots vient appuyer ce que l'on a vu avec les segments répétés. Un certain nombre de synopsis présentent le contexte/but économique et/ou technique (Ce film montre que . . . , Ce film est réalisé avec . . . , etc.).

- La thématique **Les Histoires** contenant les mêmes mots clefs que précédemment *nouvelle* est devenu *nouvel*, *raconte* est devenu *raconter* ((quadrant haut-droit de la figure 4.11). On retrouve la proximité entre *conte* et *enfant*.

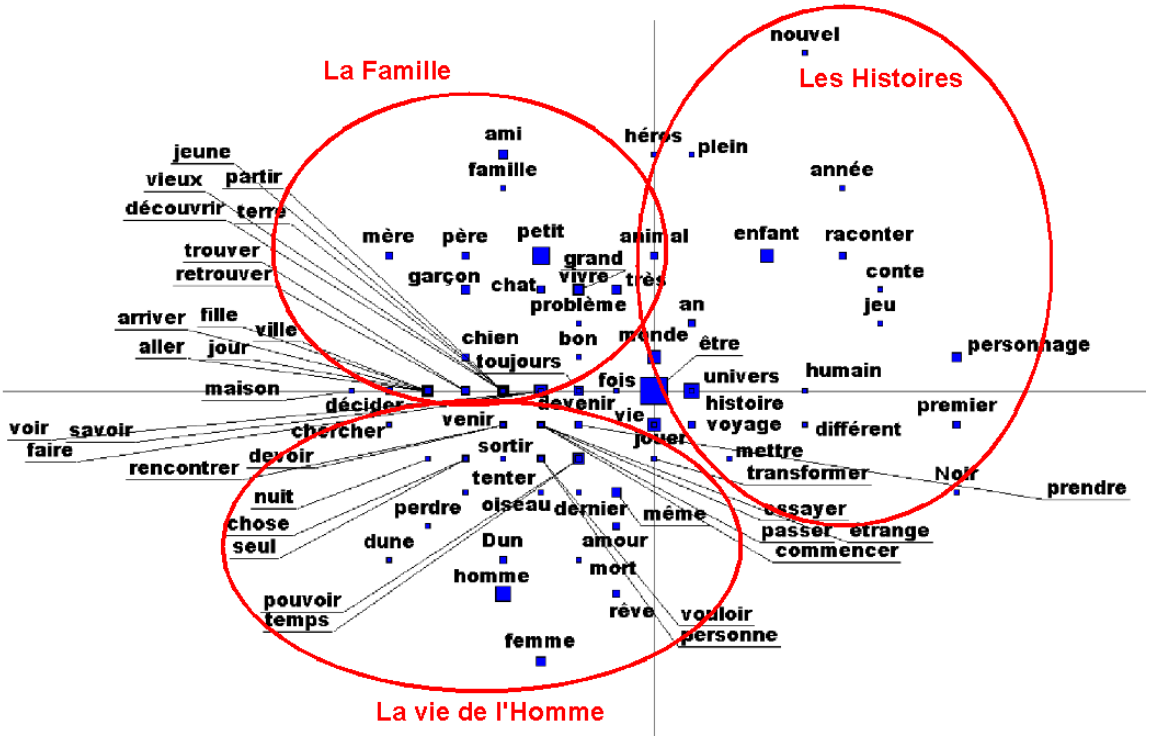



FIGURE 4.11 – Carte de l'Analyse factorielle des correspondances multiples pour la recherche de thématiques dans le corpus avec lemmatisation (zoom sur la constellation au centre du graphique)

– La thématique **La vie de l'Homme** contenant les mots clefs *"homme, femme, rencontrer, amour, mort, personne, rêve, vie"* (quadrant bas-gauche de la figure 4.11). Les termes comme travail, guerre ont disparu (nous conservons les 100 termes les plus fréquents) laissant place à un ensemble de verbes comme *perdre, tenter, sortir, pouvoir, venir, chercher* qui décrivent les **actions** de l'Homme sur son environnement. On retrouve la proximité dramatique entre *amour* et *mort*.

– La thématique **La Famille** contenant les mêmes mots clefs que précédemment (quadrant haut-gauche de la figure 4.11).

La grande différence entre ces deux analyses vient de l'ajout d'un grand nombre de verbes (les flexions grammaticales des verbes diminuaient leur fréquence dans l'analyse non lemmatisée) qui pour la plupart se situent entre les thématiques de la *famille* et de la *vie de l'homme* (*voir, savoir, faire, aller, trouver, découvrir, trouver, retrouver, etc.*).

Cette analyse lexicale des distributions globales nous a permis d'isoler les quatre thématiques du corpus qui étaient vues comme un sac de mots. Nous allons appliquer une analyse topologique afin de voir si il y a une évolution du corpus et de ses thématiques au cours du temps.

### 4.2.2 Analyse topologique

Le corpus sur lequel nous avons travaillé est en réalité un ensemble de synopsis repérés dans le temps par l'année de présentation du film au FIFA. Cette information chronologique est intéressante à exploiter pour étudier l'évolution du corpus et de ses thématiques au cours du temps. Nous allons nous intéresser maintenant à cette analyse topologique du corpus des synopsis des films d'animation.

#### 4.2.2.1 Découpage du corpus en fonction du temps

Dans un premier temps nous regroupons les synopsis par période de 10 ans. Notre corpus est divisé en classes décennales à partir de 1960 (début du festival). Avant 1960 les synopsis étant moins nombreux nous les regroupons par classes de 20 ans. La figure 4.12 montre la répartition du corpus suivant l'axe chronologique.

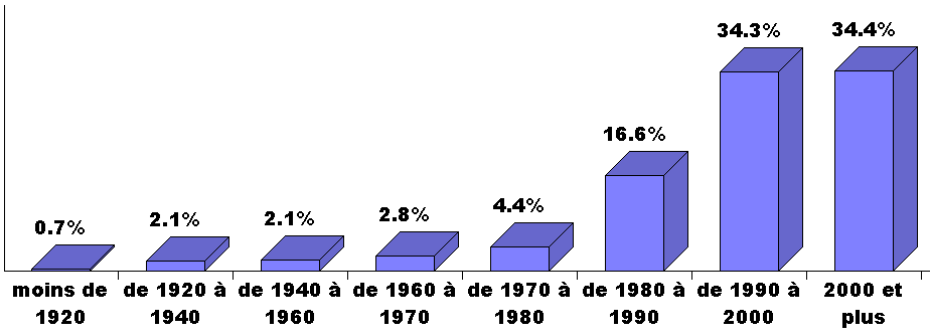



FIGURE 4.12 – Répartition des synopsis dans le temps

#### 4.2.2.2 Repérage des thématiques

Nous voulons étudier l'évolution dans le temps, des quatre thématiques retrouvées grâce à l'analyse des macrodistributions lexicales. Pour cela nous repérons dans chacun des synopsis du corpus, les lexiques des quatre thématiques. Par exemple le film THE GIRL WHO SWALLOWED BEESEST dont le synopsis est "Une jeune fille amère se lance dans un morne voyage où la magie et l'inattendu s'allient pour changer son cœur" contient les mots *fille* et *voyage* appartenant respectivement aux thématiques Famille et Histoires. Ainsi ce synopsis sera marqué comme appartenant à ces deux thématiques. Le tableau 4.3 représente la couverture de ces thématiques dans le corpus. On remarque que cette couverture est relativement bonne compte tenu de la très faible quantité de vocabulaire qui les définit. Le nombre de citations est supérieur au nombre d'observations du fait de réponses multiples.

#### 4.2.2.3 Répartition des thématiques dans le temps

Finalement, nous rapprochons le repérage des thématiques aux classes chronologiques afin de voir leur évolution au cours du temps. L'utilisation d'une table de contingence permet pour chaque sous corpus (synopsis appartenant à une classe chronologique) de repérer la microdistribution de chacune des thématiques. La carte (voir figure 4.13) issue d'une AFC

| | Nb. citation | Fréquence |
|---|---|---|
| Non réponse | 7349 | 40.5% |
| HISTOIRES | 4981 | 27.4% |
| VIE_HOMME | 3458 | 19.0% |
| FAMILLE | 2677 | 14.7% |
| FILM | 2959 | 16.3% |
| TOTAL OBS. | 18155 | |

TABLE 4.3 – Couverture des thématiques sur l'ensemble du corpus

permet de juger de la distribution des thématiques dans le temps par rapport à une situation d'indépendance.

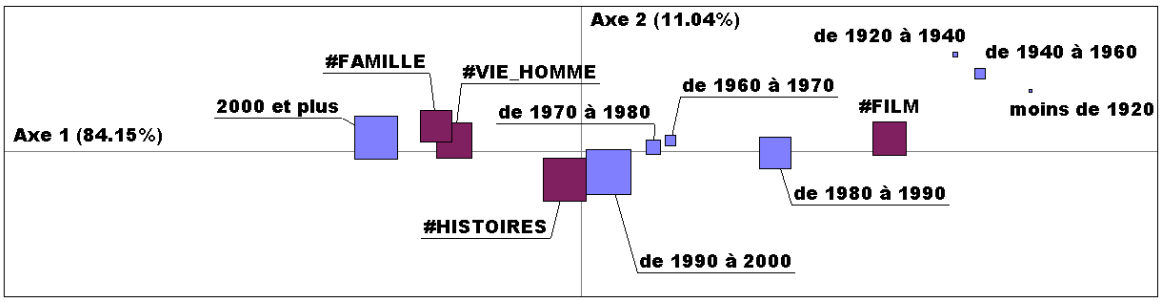



FIGURE 4.13 – Carte AFC des thématiques dans le temps

Ainsi, on remarque que la thématique du film, qui consiste le plus souvent à expliciter le contexte de production du film et/ou la technique d'animation, a tendance à être beaucoup plus présente avant 1960. Beaucoup de ces films hors-concours sont des précurseurs des techniques d'animation ce qui peut expliquer que les synopsis en font état et mettent en avant la technicité. Finalement il y a moins de place pour décrire l'histoire du film. Par contre on remarque que les autres thématiques ont une propension à se répartir de manière homogène dans le temps à partir de 1960. On peut donc conclure que les thématiques repérées par l'analyse lexicale **histoires**, **famille** et **vie de l'homme** sont probablement beaucoup plus présentes dans les films du festival que dans les films hors-concours (plus anciens).

### 4.2.3 Conclusion partielle

On a vu dans cette partie que l'analyse statistique textuelle permet d'appréhender de façon macroscopique l'information contenue dans le corpus des synopsis français. De plus cette information, on l'a vu, est stable au cours du temps quand celui-ci est réduit à la période du festival d'animation. Grâce à cette analyse globale nous avons pu isoler des caractéristiques intéressantes de ces textes comme les thématiques générales abordées dans ces films. En effet, ces films ont tendance à aborder des sujets de société c'est-à-dire qu'ils mettent en scène des **histoires** (aventure, voyage, nouvelle) centrées sur l'**Homme** (et ses actions sur son environnement), ses relations (la famille) et sa vie (l'amour, le travail, les drames, etc.).

Cependant cette information globale ou macroscopique ne rend pas compte des contenus et des détails des histoires propres à chaque film et donc à chaque synopsis. Notre objectif

étant la caractérisation de chaque film il est nécessaire de procéder à une analyse plus détaillée de chacun des synopsis. Cette analyse locale doit en plus tenir compte des spécificités de ces textes qui passent par une hétérogénéité des vocabulaires et de leur taille (une vingtaine de mots en moyenne dans un synopsis) et passent également par les thèmes abordés (mis en lumière juste avant). Nous devons mettre en place une nouvelle analyse capable d'extraire une information descriptive du synopsis, c'est la phase d'extraction d'information qui nécessite de définir au préalable le modèle de l'information que l'on cherche à extraire.

## 4.3 Modélisation d'un synopsis

L'extraction d'information consiste à extraire une information structurée à partir d'un texte. Cela nécessite de définir une structure ou un patron sur lequel l'analyse va s'appuyer afin de retrouver les informations pertinentes. Bien évidement cette structure est dépendante du contexte et de la tâche que l'on désire réaliser. Ainsi, le patron servant à la recherche d'information caractérisant les spams dans un serveur de messagerie est différent d'un patron en recherche d'information en génétique.

Dans le cas des synopsis, nous avons vu grâce à l'analyse statistique du corpus que, d'une façon globale, l'information contenue dans les synopsis pouvait être regroupée sous forme de trois thématiques principales en relation avec l'histoire racontée par le film (la quatrième - FILM - concerne une thématique connexe au film et aux techniques d'animation). Bien que ces textes soient courts ils permettent généralement de synthétiser le sujet traité par le film (qui sont eux aussi courts $\simeq$ 10 min). Ainsi à partir de l'analyse statistique du corpus et d'une expertise du domaine, un ensemble de caractéristiques communes présentes dans ces péritextes a pu être isolé. Ces caractéristiques ont pu être précisées, hiérarchisées et regroupées pour former un modèle d'information représentatif de l'information d'un synopsis. Nous cherchons donc à extraire des synopsis des informations bien précises comme les personnages, leurs actions et le contexte de ces actions (les lieux et les temps de l'histoire). Finalement dans ce modèle nous avons choisi de nous focaliser sur la recherche des actions décrites dans les textes (actions mises en lumière dans la thématique de **la vie de l'Homme**, voir figure 4.11).

### 4.3.1 Le scénario actanciel

Ces informations sont modélisées sous la forme d'un schéma spécifique que nous avons baptisé **Scénario Actanciel** (voir la figure 4.14). Ce schéma spécifique se distingue du schéma narratif et du schéma actanciel. En effet le schéma narratif [Propp *et al.*, 1970, Bremond, 1973] permet selon Propp l'analyse d'un récit et d'en extraire la structure en s'appuyant sur les éléments suivants :

- la **situation initiale**, (ou état initial).
- l'**élément déclencheur** qui modifie la situation initiale.
- les **péripéties** (toutes les actions) entreprises par le(s) héro(s) pour atteindre son (leur) but.

– l'**élément de résolution** (dénouement) qui conduit à la situation finale.

– la **situation finale** qui est la fin du récit.

Le schéma actanciel [Greimas, 1966] explicite les relations entre les actants qui permettent le récit. On retrouve dans ce dernier les éléments suivants :

– Le **destinateur** est le mandateur qui pousse le héros à agir, celui qui l'envoie en mission.

– Le **sujet** (ou héros) est celui qui accomplit l'action, celui qui effectue la quête.

– L'**objet** est ce que cherche le sujet ou ce qu'il doit accomplir.

– Le **destinataire** est le bénéficiaire de l'action du sujet.

– L'**opposant** qui nuit au sujet et l'empêche d'agir.

– L'**adjuvant** est la personne (ou l'objet) qui vient en aide au sujet, lui permettant de surmonter les épreuves auxquelles il se trouve confronté.

Dans le cas des synopsis ces schémas sont inadaptés car beaucoup trop complexes. En effet les synopsis ne sont pas toujours des textes narratifs et ils sont bien trop courts pour apporter la richesse informationnelle nécessaire à l'instanciation de ces modèles.
Ainsi nous proposons un schéma beaucoup plus simple et proche de l'information contenue dans les textes dont nous disposons.

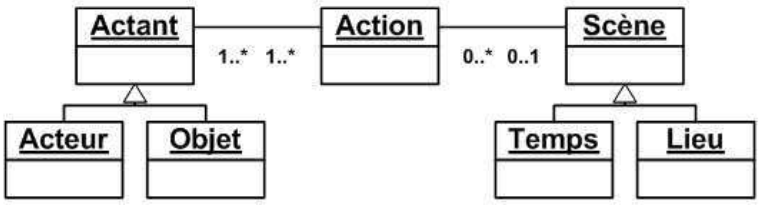



FIGURE 4.14 – Le scénario actanciel sous forme d'un schéma UML

Notre **scénario actanciel** ou modèle d'information (figure 4.14) nous permet de nous focaliser sur l'**action**. L'**action** (ou procès[8] du verbe) fait intervenir des **actants** (sujets et compléments d'objets du verbe) qui sont des **acteurs** (ceux qui participent à l'action) ou des **objets** (ceux qui subissent ou permettent l'action). Cette action se déroule dans une **scène** qui est la description du contexte temporel et/ou locatif (compléments circonstanciels de lieux et/ou de temps). Ce modèle nous permet donc de capturer l'information de type action contenue dans les quelques lignes d'un synopsis.

8. On dit d'un verbe qu'il indique un procès quand il exprime une action réalisée par le sujet, par opposition notamment aux verbes exprimant un état ou un résultat. *Dictionnaire de linguistique, Larousse, 1991*

### 4.3.2 Exemple

Voici un exemple pour expliciter notre approche. Prenons le synopsis du film **Bus Stop** :

*« Un caribou et un Canadien se disputent un peu de place sur un banc en attendant l'autobus. »*

nous avons :

- 3 **Actants** : dont 2 **Acteurs** (*un caribou* et *un Canadien*) et un **Objet** (*la place*)
- 1 **Action** : *se disputer*
- 1 **Scène** -lieux : *banc* -temps : *en attendant l'autobus*

ce qui donne le scénario actanciel de la figure 4.15.

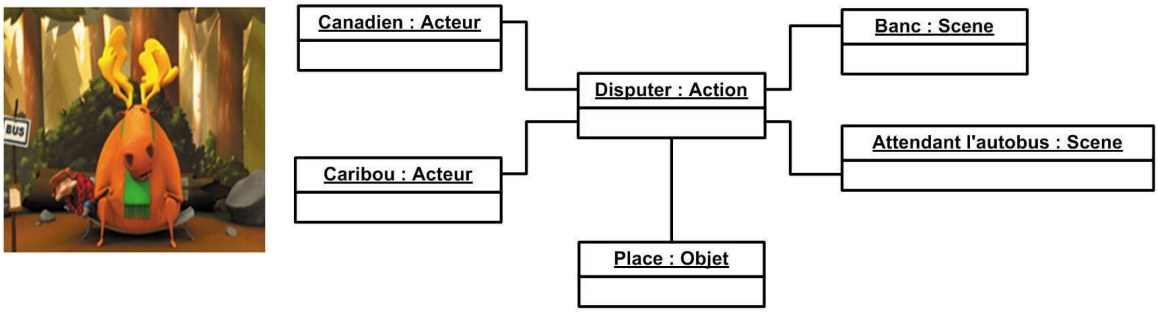



FIGURE 4.15 – Le scénario actanciel (sous forme d'un schéma UML) du synopsis du film **Bus Stop**

## 4.4 L'Extraction d'Information

Le but de cette étape est d'extraire automatiquement à partir d'une ressource textuelle réduite, en l'occurrence un synopsis, une information structurée selon notre **scénario actanciel**. En réalité l'Extraction d'Information (EI) en traitement de la langue est une sorte de système de recherche d'information qui consiste à retrouver une information structurée à l'intérieur du texte. Préalablement, il faut être capable d'expliciter et structurer l'information que l'on désire retrouver dans les textes à analyser. L'EI est donc une approche différente de l'Extraction de Connaissances qui consiste à analyser un corpus volumineux sans aucune connaissance *a priori* sur le(s) texte(s) afin d'en extraire une information.

### 4.4.1 Les étapes

La chaîne de traitements permettant de réaliser cette tâche automatiquement est une chaîne assez classique en EI [Kosseim et Lapalme, 1998, Muslea, 1999, Cunningham, 2005, Sarawagi, 2008]. Le schéma que nous avons adopté (figure 4.16) est composé d'un ensemble

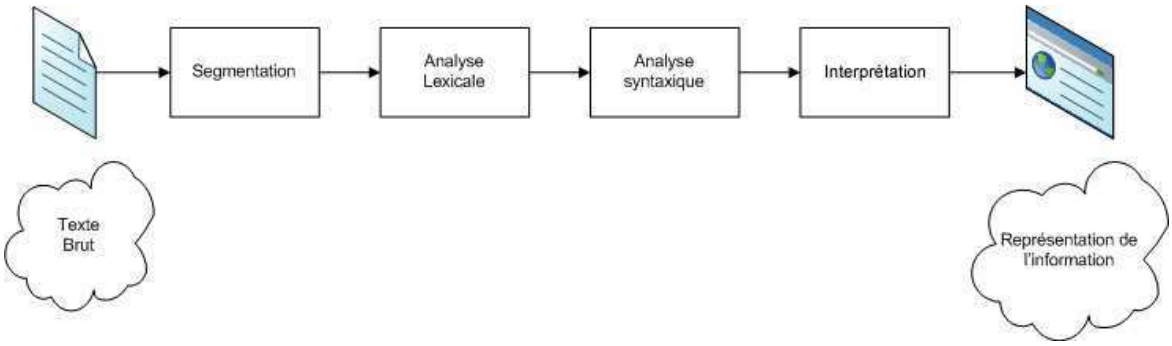



FIGURE 4.16 – Diagramme des différentes étapes de l'extraction d'information à partir d'un synopsis

d'étapes qui permettent de passer du texte non ou peu structuré vers sa représentation structurée.

Nous allons détailler brièvement l'ensemble de ces étapes :

- La **segmentation** de texte est une phase nécessaire pour un très grand nombre d'applications en traitement automatique du langage. Elle permet le découpage du texte en unités homogènes et interprétables comme les mots, les phrases ou les paragraphes. La segmentation s'appuie le plus souvent sur les signes de ponctuation qui constituent des marques pivot pour identifier les phrases [Mourad, 1999].

- L'**Analyse lexicale** vise à transformer un ensemble de caractères en un lexème (mot). Cette analyse se base sur un étiqueteur lexical qui recherche la catégorie lexicale de chaque mot (article, nom, verbe, déterminant, etc.) en se basant sur des lexiques, mais également sur des outils de désambiguïsation sémantique. Par exemple dans la phrase « *je plante une plante verte* » le mot *plante* peut être un nom ou un verbe.

- L'**Analyse syntaxique** consiste à exhiber la structure d'un texte construit en utilisant un ensemble de règles de syntaxe formant une grammaire formelle. Cette analyse donne alors précisément la façon dont les règles de syntaxe sont combinées dans le texte. Connaître la structure syntaxique d'un énoncé permet d'expliciter les relations de dépendance (par exemple entre sujet et objet) entre les différents lexèmes, puis de construire une représentation du sens de cet énoncé.

- La tâche d'**Interprétation** consiste à ajouter de l'information sémantique à l'analyse syntaxique afin de permettre l'application de règles d'extraction pour instancier le modèle de représentation de l'information qui, dans notre cas, est le *scénario actanciel*.

La tâche d'**Interprétation** qui consiste en une analyse sémantique s'appuie généralement sur 5 grandes étapes définies par la Message Understanding Conference (MUC) [Grishman, 1996] que l'on retrouve dans la plupart des systèmes d'EI [Cunningham, 2005].

**Named Entity (NE) Recognition** , "*la reconnaissance des entités nommées*", comme les personnes et les organisations mais également les lieux, les expressions temporelles et certains types d'expressions numériques.

**Coreference resolution (CO)** ,"*la résolution des coréférences*", consiste en l'identification des mots ou segments qui se réfèrent au même objet. Par exemple l'anaphore est un type de coréférence. Dans la phrase « *Jean n'avait pas de stylo : je lui ai prêté le mien* », le pronom possessif « le mien » est une anaphore dont l'antécédent est le nom « stylo ».

**Template Element (TE) construction** , "*la résolution des descriptions*", consiste en la construction des tables d'éléments dans lesquelles sont ajoutées aux entités reconnues par l'étape de NE les informations descriptives contenues dans le texte.

**Template Relation (TR) construction** , "*la résolution des relations*", consiste en l'identification des relations entre les entités trouvées dans l'étape de TE. Cela peut être par exemple des relations de type *employé* (entre une personne et une entreprise).

**Scenario Template (ST) production** ,"*l'instanciation du modèle d'information*", c'est la sortie du système. Il remplit le modèle de représentation de l'information prédéfinie par l'utilisateur en liant ensemble les entités extraites de l'étape du TE avec les relations extraites du TR.

Afin d'illustrer ces différentes étapes prenons l'exemple suivant : « *La fusée rouge brillant a été mise à feu mardi. C'est l'invention du Dr Head. Il fait partie du personnel scientifique de Rockets Inc.* ».

- L'étape du NE extrait les entités suivantes *fusée*, *mardi*, *Dr. Head* et *Rockets Inc.*
- L'étape du CO trouve que *c'est* se réfère à *la fusée* et que *Il* se réfère au *Dr. Head.*
- L'étape du TE découvre que la *fusée* est *rouge brillant* et qu'elle est l'*invention de Head.*
- L'étape du TR découvre que le *Dr. Head travaille pour l'entreprise Rockets Inc.*
- L'étape du ST découvre que l'événement présenté ici est le lancement d'une fusée et que les différentes entités y sont impliquées.

Nous venons de voir qu'une analyse sémantique du texte permet d'identifier les informations pertinentes afin d'instancier le modèle d'information. Dans le cadre de l'analyse des synopsis, nous allons nous intéresser aux premières étapes qui consistent à analyser le lexique et la syntaxe du texte. Puis nous verrons comment la tâche d'interprétation est réalisée pour instancier le *scénario actanciel.*

### 4.4.2 L'analyse syntaxique

Un premier travail, préliminaire à l'analyse sémantique, est l'analyse **syntaxique** dénommée aussi analyse **linguistique** qui consiste à étudier les dépendances syntaxiques entre

unités c'est-à-dire d'exhiber la structure d'un texte écrit dans une langue naturelle. Un analyseur syntaxique (parser, en anglais) est un programme informatique qui réalise cette tâche automatiquement. La structure révélée par l'analyse est en fait la structure grammaticale qui donne alors précisément la façon dont les règles de syntaxe sont combinées dans le texte. Par exemple les unités interprétables pour qualifier un objet ou un concept du monde réel sont en général les groupes nominaux [Haddad et Chevallet, 2003, Turenne, 2001].

De plus, l'instanciation automatique du scénario actanciel est fortement dépendante de cette analyse syntaxique. En effet, le sujet d'un verbe d'action sera toujours un **Actant**, les compléments circonstanciels de lieu et de temps seront quant à eux les éléments de la **Scène**. Contrairement aux approches classiques, ce n'est pas l'étape de reconnaissance des entités nommées (NE) qui permettra de retrouver à coup sûr les *acteurs* de notre *scénario actanciel*. En effet, dans un texte classique les acteurs sont des êtres humains que l'on retrouvera grâce à la reconnaissance des entités nommées. Dans le cas des synopsis, cet *a priori* est remis en question. Par exemple dans la phrase : *"Les trois petits cochons s'enfuient et se réfugient dans la maison de briques"*, la reconnaissance des entités nommées (NE) ne permet pas de retrouver les actants du film (un cochon est un animal et non une personne) c'est principalement la syntaxe et la reconnaissance des règles grammaticales (ici le lien entre le sujet et le verbe d'action) qui conduiront à l'instanciation du modèle.

Pour réaliser cette analyse nous avons décidé d'utiliser les synopsis anglais car l'ensemble des outils de Traitement Automatique de la Langue (TAL) disponibles et modifiables (en "open source") sont difficiles à trouver pour le français. Nous utilisons donc l'analyseur Link Grammar (LG) qui est un *"parser"* de la langue anglaise implémentant déjà une segmentation et une analyse lexicale. De plus ce *"parser"* est régulièrement amélioré [9] et on le retrouve dans de nombreux travaux [Madhyastha *et al.*, 2003, Pyysalo *et al.*, 2006, Curtis *et al.*, 2009, Hakenberg *et al.*, 2009].

#### 4.4.2.1 Un analyseur syntaxique : Link Grammar

LG est un parser pour la langue anglaise [Sleator et Temperley, 1991] basé sur le modèle introduit par Lucien Tesnière des grammaires de dépendance (Dependency Grammar) dont le principe est qu'un mot dépend d'un autre. Dans une phrase, les mots ne font pas que se suivre, ils entretiennent également des relations. La principale caractéristique de ces grammaires est que la structure d'une phrase est vue comme un mot appelé tête (verbe), auquel sont attachés des modificateurs (par exemple les sujets, les compléments). Les modificateurs peuvent à leur tour posséder des modificateurs et la structure d'une phrase devient une arborescence (figure 4.17) [Aubin, 2002].

Plus spécifiquement LG utilise un ensemble de mots définis dans un dictionnaire et contraints par des règles de combinaison (linking requirement). Une phrase grammaticalement correcte est donc une séquence de mots telle qu'il existe un chemin qui permet de connecter tous les mots entre eux. La figure 4.18 montre un tel dictionnaire pour les mots *a, the, cat, snake, Mary, ran,* et *chased*.

Chaque mot a un ou plusieurs connecteurs qui peuvent s'enficher les uns dans les autres (sous réserve d'être compatibles). Ainsi le mot *cat* à besoin de saturer ses connecteurs (point noir figure 4.18). Il doit donc mettre en jeu un connecteur de type D (connexion entre un déterminant et un nom) à sa gauche **ET** l'un des deux connecteurs : de type O (connexion

9. http://www.abisource.com/projects/link-grammar

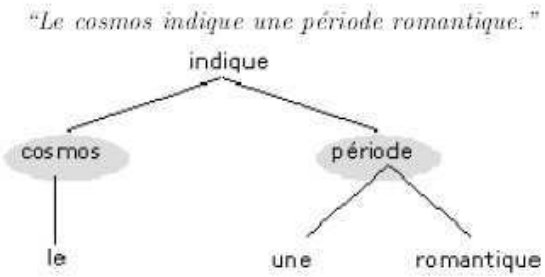



FIGURE 4.17 – Arbre des dépendances syntaxiques

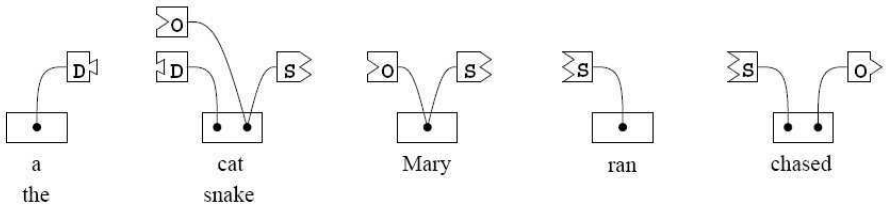



FIGURE 4.18 – Extrait du dictionnaire de Link Grammar

entre un verbe et un objet) à sa gauche **OU** un connecteur de type S (connexion entre un nom sujet et un verbe) à sa droite. Le diagramme 4.19 montre comment l'ensemble des mots de la phrase *The cat chased a snake* peuvent créer un ensemble de liens (« linkage ») corrects (*Correct* : signifiant que les liens ne se croisent pas et qu'ils suffisent à connecter tous les mots ensemble).

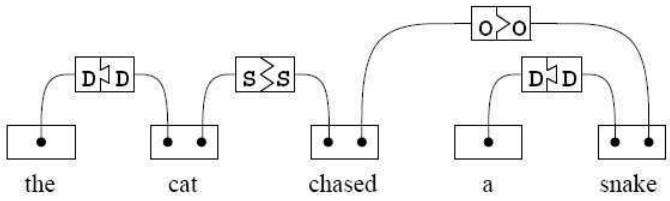



FIGURE 4.19 – Diagramme de liaisons correctes

Il est facile de voir que les phrases *Mary chased the cat* et *The cat ran* sont aussi des phrases grammaticalement correctes. Par contre comme on peut le voir sur la figure 4.20 la phrase *the Mary chased cat* est incorrecte.

Ainsi l'analyse de la structure grammaticale d'une phrase est déterminée par les connecteurs mis en jeu lors de la création des liens (« linkage ») de cette dernière, chaque connecteur ayant un rôle syntaxique particulier (par exemple le connecteur S correspond au rôle de **sujet** par rapport au **verbe**). Par contre une des difficultés rencontrée avec LG est qu'il crée plusieurs schémas possibles pour une même phrase, cela pouvant aller jusqu'à plusieurs centaines de schémas de liaisons (« linkages ») pour une phrase contenant une dizaine de mots. Dans [Kakkonen, 2008, Pyysalo *et al.*, 2006, Molla et Hutchinson, 2003] les auteurs n'utilisent que le premier résultat retourné par l'algorithme qui est le schéma le plus plausible. Or ne prendre

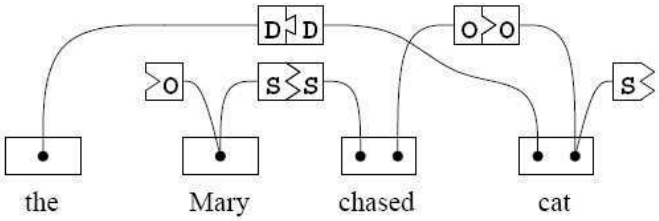



FIGURE 4.20 – Diagramme de liaisons incorrectes

qu'un schéma (« linkage ») ne permet pas d'obtenir la structure syntaxique complète de la phrase et notamment lorsqu'il y a plusieurs sujets. Dans la phrase *John and Matt run on the road* ne prendre que le premier « linkage » fait que l'on perd un des 2 sujets (figure 4.21). Afin de remédier à ce problème nous avons programmé un module statistique qui se focalise sur le verbe et ses relations. Ainsi tous les schémas retournés par LG sont pris en compte et chaque mot se voit attribué la catégorie où il apparaît le plus de fois.

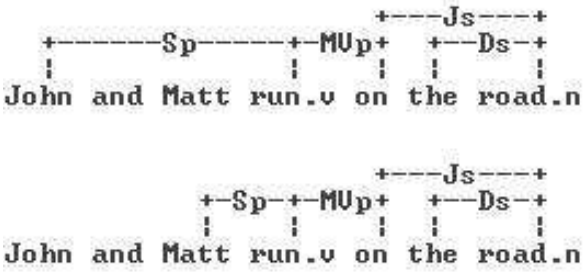



FIGURE 4.21 – Résultats de Link Grammar quand il y a 2 sujets

Dans l'exemple de la figure 4.21 le module regarde (parmi les 2 schémas retournés par LG), pour chaque mot, les liens qui lui sont attribués et plus particulièrement les liens du type sujet (**S**), verbe (**V**), objet (**O**) et complément adverbial (**A**). On crée 4 tables (une pour chaque type) de contingence dans lesquelles on note le nombre de fois (parmi tous les schémas retournés) où le mot a été catégorisé sujet, verbe, objet ou complément adverbial. On obtient pour cet exemple les tables suivantes (chaque case correspond à un mot de la phrase)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **S** | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| **O** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **A** | 0 | 0 | 0 | 0 | 2 | 2 | 2 |

A partir de ce tableau il est facile de calculer des paramètres statistiques et d'affecter chaque mot à sa catégorie syntaxique la plus probable. Dans cet exemple les sujets sont *John* and *Matt*, le verbe est *run*, il n y a pas de complément d'objet et le complément adverbial est *on the road*. Finalement ce module permet de retrouver les 4 groupes ou éléments syntaxiques importants pour instancier le scénario actanciel. Voir l'annexe D.1 pour différents exemples sur des phrases issues des synopsis.

#### 4.4.2.2 Tests et résultats

Une évaluation de cet outil est réalisée à partir d'une vérité terrain construite à la main et constituée d'une trentaine de synopsis. Le tableau 4.5 présente une synthèse des résultats obtenus pour chacune des catégories grammaticales sous la forme d'une mesure de précision et d'une mesure de rappel.

| | **Sujet** | **Verbe** | **Objet** | **Adverbial** |
|---|---|---|---|---|
| **Précision** | 0.93 | 0.98 | 0.95 | 0.96 |
| **Rappel** | 0.91 | 0.96 | 0.93 | 0.88 |

TABLE 4.4 – Résultats de l'analyse syntaxique sous forme de Précision et de Rappel

On remarque à travers ces résultats que l'algorithme retrouve avec une très bonne précision les verbes et les groupes verbaux ($\simeq$ 98%). Cependant ils ne sont pas toujours tous retrouvés (rappel de 96%). Cette remarque est également valable pour les sujets et groupes sujets. Ceci vient du fait que la recherche des groupes sujets est fortement liée à la recherche des verbes. Bien que les synopsis soient très souvent composés d'un sujet et d'un verbe clairement identifiables, il y a cependant des phrases très complexes qui sont par exemple composées de plusieurs verbes n'ayant qu'un seul sujet. Dans cette situation le rappel des verbes chute car l'algorithme tente de retrouver des couples sujet - verbe suivis d'un complément du verbe. Il existe des situations (comme dans la phrase suivante par exemple) où ces couples sont difficilement identifiables :

*"The experiences of a paper character crumpled, rolled up, distorted, thrown away, shaken about, ending up by an encounter which is both sweet and light"*

l'algorithme détecte 166 schémas (« linkage ») possibles et détecte bien les 7 verbes puis retourne finalement trois structures Sujet-Verbe-Objet-Adverbial :

**Sujet :** "the **experiences** of a paper character"

**Verbe :** "crumpled"

**Objet :** $\emptyset$

**Adverbial :** "rolled up"

Dans cette structure le sujet et le premier verbe sont correctement retrouvés alors que le deuxième verbe est interprété par l'algorithme comme un complément adverbial du premier verbe car celui-ci vient juste après. Puis il retourne la deuxième structure suivante :

**Sujet :** "the **experiences** of a paper character"

**Verbe :** "distorted"

**Objet :** $\emptyset$

**Adverbial :** "crumpled", "rolled", "thrown"

L'algorithme retourne une deuxième structure identique à la première où les verbes sont vus comme des compléments du verbe "distorted". Ceci vient du fait que l'algorithme tente de retrouver des « linkage » où chaque mot met en jeu des connecteurs qui lui sont propres.

Or les verbes ne sont généralement pas connectés les uns aux autres ce qui conduit à l'interprétation erronée précédente. Finalement seule la dernière structure (plus conventionnelle) est retrouvée :

**Sujet :** "an **encounter**"

**Verbe :** "is"

**Objet :** $\emptyset$

**Adverbial :** "sweet ", "light"

Les compléments d'objets et compléments adverbiaux sont beaucoup plus difficiles à extraire de ces textes, même par un opérateur humain en raison de la complexité syntaxique de certains synopsis. De plus, la distinction entre les différents types de compléments est un processus complexe qui se fait bien souvent sur la base de la compréhension sémantique de la phrase. De plus, dans certaines situations (voir l'exemple ci dessus) des verbes peuvent être interprétés comme des compléments adverbiaux ce qui augmente le nombre de mauvaises détections faisant ainsi chuter la précision. Finalement, nous sommes arrivés à la conclusion assez naturelle que cet algorithme donne de bon résultats lorsque les phrases sont courtes et bien formées. Par contre les résultats sont dégradés lorsque les phrases deviennent longues et complexes. De plus les résultats de ces analyses restent satisfaisants, ce qui permet d'aborder la phase d'instanciation du scénario actanciel.

### 4.4.3 La tâche d'Interprétation

La tâche d'Interprétation (voir figure 4.22) a pour but d'instancier le modèle d'information. Pour cela cette tâche s'appuie sur l'analyse syntaxique et sur la reconnaissance sémantique des éléments du texte. Dans ce dernier cas, il est nécessite d'avoir des connaissances pour reconnaitre et construire le modèle d'information. Ainsi, l'utilisation de ressources sémantiques extérieures comme le thésaurus WordNet permet d'avoir un référentiel assez complet et général pour pouvoir mettre en place cette tâche.

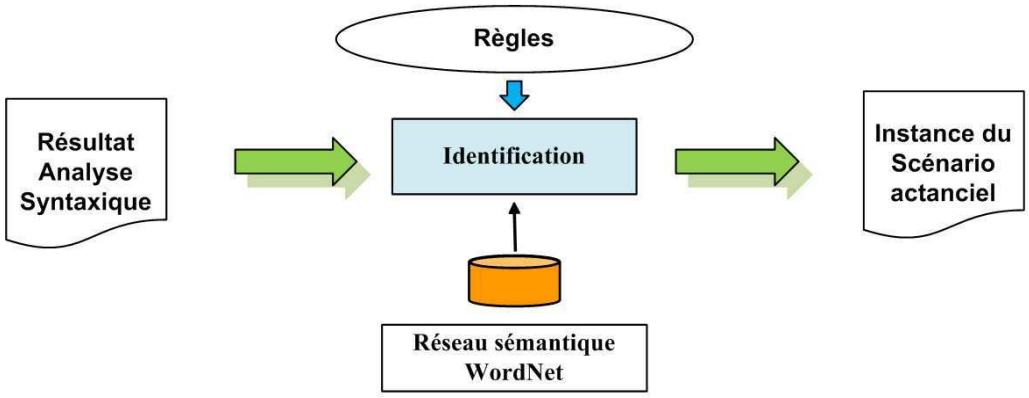



FIGURE 4.22 – Instanciation du scénario actanciel

L'instanciation du scénario actanciel (voir figure 4.22) se fait à partir des trois ressources suivantes :

– L'information syntaxique : Cette information permet de connaitre les liens syntaxiques qui lient les différent éléments de la phrase. Or nous l'avons vu précédemment, les

informations présentes dans le scénario actanciel sont fortement liées à la structure syntaxique de la phrase. Par exemple, l'**Action** du scénario actanciel sera toujours un verbe dont le sujet sera toujours un **Actant**.

- L'information sémantique : Cette information apportée par le réseau sémantique WordNet, permet de caractériser sémantiquement les éléments constituant la phrase. Ce réseau de concepts nous permet de filtrer les éléments de la phrase suivant leur catégorie conceptuelle.

- Les règles d'instanciation : Un ensemble de règles permet, à partir des informations de syntaxe et des informations conceptuelles, d'attribuer aux éléments de la phrase une place et un rôle dans le scénario actanciel.

Puisque l'analyse syntaxique a été vue précédemment nous allons maintenant détailler l'utilisation du thésaurus WordNet ainsi que les règles d'instanciation.

#### 4.4.3.1 Un thésaurus : WordNet

WordNet [Miller, 1995] est une base de données lexicales ($\simeq 150,000$ mots) développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Les noms, verbes, adjectifs et adverbes sont organisés en ensembles de synonymes (synsets $\simeq 115,000$), représentant un groupe de mots interchangeables, dénotant un sens ou un usage particulier. WordNet [10] répertorie ainsi une grande variété de relations sémantiques permettant d'organiser le sens des mots (et donc par extension les mots eux-mêmes). Il regroupe de nombreux mots et concepts de la langue anglaise en restant indépendant d'un domaine. Ceci est important pour nous car les synopsis n'utilisent pas un vocabulaire spécialisé et sont très hétérogènes.

On peut définir de manière succincte ces relations comme suit (il existe d'autres relations plus spécifiques aux verbes) :

- Relation **Synonymie** : le synset (synonym set), représente un ensemble de mots qui sont interchangeables dans un contexte donné.

- Relation **Hyperonymie** : c'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques. Y est un hyperonyme (hypernyme) de X si X est une sorte de (kind of) Y.

- Relation **Hyponymie** : c'est le terme utilisé pour désigner une classe spécifique d'une classe générique (relation inverse de Hypernymie). X est un hyponyme de Y si X est un type de (kind of) Y.

- Relation **Holonymie** : c'est le terme utilisé pour désigner une classe constituée d'autres classes (méronymes). Y est un holonyme de X si X est une partie de (is a part of ) Y.

10. Nous avons réalisé une version pour Windows qui est disponible à http://sourceforge.net/projects/wordnet30forwin

– Relation **Méronymie** : c'est le terme utilisé pour désigner une classe/partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de l'Holonymie). X est un méronyme de Y si X est une partie de Y. exemple : *avion* a comme meronymes *porte*, *moteur* ; *moteur* a comme meronymes *hélice*, *réacteur*.

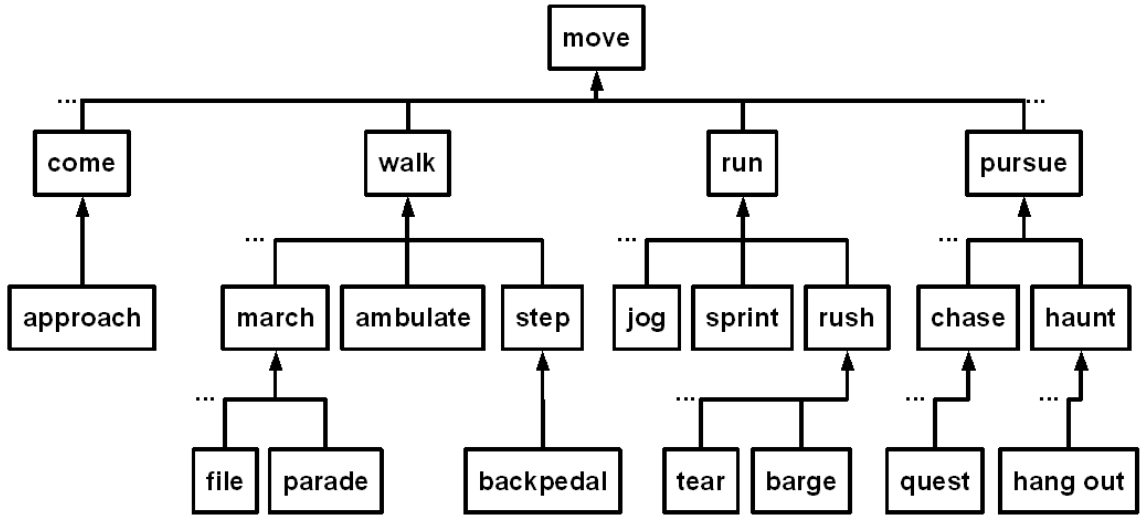



FIGURE 4.23 – Extrait d'une partie de la hiérarchie des hyponymes de "move" sous WordNet

La figure 4.23 présente une sous hiérarchie de WordNet correspondant au concept "move". A partir de cette hiérarchie il est facile de retrouver les verbes de mouvement. En effet, pour un élément donné il suffit de parcourir la hiérarchie de ses hyperonymes afin de vérifier qu'il est un hyponyme du concept général *"move"*. Cette ressource et ce mécanisme de parcours du réseau des hyperonymes nous permet de caractériser sémantiquement les éléments de la phrase à partir d'un ensemble de concepts généraux de base liés au scénario actanciel. Nous avons défini trois super-concepts liés aux classes du scénario actanciel qui contiennent chacun des concepts de base de WordNet :

– L'**Action**. Un élément $X$ appartient à ce super-concept si il a au moins dans son réseau d'hyperonymes un des concepts de base de WordNet inclus dans le super-concept Action. Nous définissons ce super-concept comme l'ensemble des concepts de base suivant :

$$\textbf{Action} = \{travel, go, move, locomote, change\}$$

$$\begin{array}{rll} X \in Action\ Si : & travel \in Hyperonyme(X) & OU \\ & go \in Hyperonyme(X) & OU \\ & move \in Hyperonyme(X) & OU \\ & locomote \in Hyperonyme(X) & OU \\ & change \in Hyperonyme(X) & \end{array}$$

– Le **Lieu**. Un élément $X$ appartient à ce super-concept si il a au moins dans son réseau d'hyperonymes un des concepts de base de WordNet inclus dans le super-concept Lieu. Nous définissons ce super-concept comme l'ensemble des concepts de base suivant :

$$\textbf{Lieu} = \{area, location, place, land, construction\}$$

$$\begin{array}{lll} X \in Lieu\ Si : & area \in Hyperonyme(X) & OU \\ & location \in Hyperonyme(X) & OU \\ & place \in Hyperonyme(X) & OU \\ & land \in Hyperonyme(X) & OU \\ & construction \in Hyperonyme(X) & \end{array}$$

– Le **Temps**. Un élément $X$ appartient à ce super-concept si il a au moins dans son réseau d'hyperonymes un des concepts de base de WordNet inclus dans le super-concept Temps. Nous définissons ce super-concept comme l'ensemble des concepts de base suivant :

$$\textbf{Temps} = \{event, time\ period, time\ unit\}$$

$$\begin{array}{lll} X \in Temps\ Si : & event \in Hyperonyme(X) & OU \\ & time\ period \in Hyperonyme(X) & OU \\ & time\ unit \in Hyperonyme(X) & \end{array}$$

Les super-concepts liés aux actants et notamment aux acteurs n'ont pas été développés ci dessus. En effet, comme nous l'avons déjà dit, les connaissances *a priori* du monde réel ne s'appliquent pratiquement plus dans le monde de l'animation. Dans notre monde réel un *acteur* est un être vivant qui est à l'origine de l'*action*, c'est donc une entité qui est animée et vivante. Dans WordNet cela reviendrait à définir le super-concept Acteur qui contiendrait un des concepts de base comme "person" ou "living thing". Cependant cette caractéristique des actants n'est plus valable dans le cadre de l'animation puisque par définition un film d'animation est une œuvre dans laquelle l'auteur donne l'illusion de la vie à des objets qui par nature sont inertes. Par exemple Bob l'éponge n'est pas défini comme une personne dans WordNet car le mot *"sponge"* est défini comme un *"absorbent material"*. Dans le film *The Inspector Goes Back Home* le synopsis décrit l'histoire suivante : *« An inspector walks the streets of the town, carefully trying to avoid danger. At home he washes his feet and reads the newspaper. Suddenly he notices a finger-print starting to run away. The inspector stalks its trail all over town. In the end he realises that he has been pursuing the print of his own finger. »*. Dans ce synopsis l'empreinte digitale ("finger-print") est personnalisée puisqu'elle réalise une action qui est *courir*. Finalement cette entité qui par définition est inanimée (dans WordNet "finger-print" appartient au concept "abstraction, abstract entity") doit apparaître comme un Actant dans le scénario actanciel. Par conséquent, la seule restriction que nous faisons pour définir un actant provient de la catégorie lexicale auquel appartient l'élément que l'on cherche à catégoriser. En effet, un actant doit obligatoirement être un **nom**, nous excluons donc les adjectifs et les adverbes.

#### 4.4.3.2 Les règles d'instanciation

Lorsque l'on dispose de la syntaxe et de la catégorie sémantique des éléments de la phrase nous pouvons grâce à l'utilisation de règles, instancier les différentes classes du scénario actanciel. En effet, ces règles permettent de relier les connaissances syntaxiques et conceptuelles aux catégories du scénario actanciel. Nous présentons ici l'ensemble de ces règles pour chacune des catégories du scénario actanciel :

– L'**Action**. Cet élément fondamental du scénario actanciel est lié à un verbe conjugué [11] et est défini comme ceci :

11. le verbe est lié à au moins un sujet et à au moins un complément (d'objet direct/indirect ou adverbial)

*l'élément X est une Action Si :* $X$ *est un* ***verbe*** *conjugué* *ET* $X \in Action$

– Les **Acteurs**. Ces éléments sont les initiateurs de l'action et sont définis comme ceci :

*X est un Acteur Si :*
*X est un :* ***nom*** *ou* ***nom propre*** *ou* ***pronom*** *ET*
*X est* ***sujet*** *du verbe d'action*

– Les **Objets**. Ces éléments sont spectateurs de l'action et sont définis comme ceci :

*X est un Objet Si :*
*X est un :* ***nom*** *ou* ***nom propre*** *ou* ***pronom*** *ET*
*X est complément d'****objet*** *du verbe d'action*

– Les **Scènes Locatives**. Ces éléments sont le contexte locatif où se produit l'action et sont définis comme ceci :

*X est un Lieu Si :*
*X est complément* ***adverbial*** *du verbe d'action* *ET*
$X \in Lieu$

– Les **Scènes Temporelles**. Ces éléments sont le contexte temporel où se produit l'action et sont définis comme ceci :

*X est un Temps Si :*
*X est complément* ***adverbial*** *du verbe d'action* *ET*
$X \in Temps$

Voir l'annexe D.1 pour différents exemples sur des phrases issues des synopsis. Finalement la différence entre les Acteurs et les Objets est liée à leur relation syntaxique par rapport au verbe de l'action.

#### 4.4.3.3 Tests et résultats

Une évaluation de cet outil est réalisée à partir d'une vérité terrain construite à la main et constituée d'une trentaine de synopsis. Le tableau présente les résultats pour chacune des catégories grammaticales sous forme de précision et de rappel.

On remarque à travers ces résultats que l'algorithme retrouve avec une très bonne précision les actions. Ce résultat vient du fait que l'analyse syntaxique retrouve généralement assez bien les verbes de la phrase (le rappel de l'analyse syntaxique est de 96%). De plus, lorsque l'on filtre ceux-ci avec WordNet on élimine les éléments qui auraient pu être classés

| | Acteur | Objet | Action | Scène lieu | Scène temps |
|---|---|---|---|---|---|
| **Précision** | 0.94 | 0.97 | 1 | 0.91 | 0.97 |
| **Rappel** | 0.91 | 0.95 | 1 | 0.86 | 0.91 |

TABLE 4.5 – Résultats de l'instanciation automatique du Scénario Actanciel

abusivement comme des verbes. Enfin le rappel est proche de 100% car les verbes non retrouvés par l'analyse syntaxique ne sont quasiment jamais des verbes d'action. Cela vient du fait que LG retrouve très bien les schémas Sujet+Verbe+Complément du verbe qui sont les schémas liés au scénario actanciel et donc aux verbes d'action.
Les actants quant à eux obtiennent des résultats identiques (en rappel) à ceux de l'analyse syntaxique des groupes sujets et objets, ce qui est cohérent avec les règles d'instanciation. En effet, l'utilisation du thésaurus permet simplement de filtrer les résultats de l'analyse syntaxique et donc d'améliorer les résultats de précision. Les résultats liés à la recherche de la scène sont les moins bons. Ceci est dû à la grande variété des compléments adverbiaux retrouvés par LG. En effet, ils peuvent être de nombreuses natures comme des compléments circonstanciels de lieu ou de temps (ceux qui nous intéressent ici) mais également des compléments circonstanciels de manière, de cause , de conséquence, etc. C'est l'utilisation de WordNet qui permet de distinguer les lieux des temps. Cependant notre approche ne tient pas compte de la polysémie [12] des termes rencontrés. Ainsi dans WordNet de nombreux termes peuvent, suivant leurs sens, appartenir au super-concept de lieu et/ou de temps.

## 4.5 Analyse thématique

Nous venons de voir qu'il est possible à partir du synopsis d'obtenir un modèle et une description détaillée de l'action mise en œuvre dans les films d'animation. Cependant cette information est très souvent liée à une description locale de ce qui se passe dans la séquence vidéo, elle est donc plus difficile à exploiter pour caractériser globalement le film. Dans cet objectif, nous envisageons dans nos travaux la caractérisation globale de la séquence d'animation à travers la sensation ou atmosphère dégagée par le film. Cette caractéristique qui correspond généralement au genre du film est une information importante pour caractériser l'œuvre dans son ensemble. De plus, cette caractérisation globale à travers le **genre du film d'animation** est motivée par le fait que ce champ est très souvent mal ou peu renseigné dans les fiches d'inscription au festival. Dans notre contexte, cette description générale de la séquence est envisagée par l'analyse du synopsis et du vocabulaire utilisé pour décrire le film. En effet les champs lexicaux des vocabulaires utilisés dans les textes sont les marqueurs d'une volonté de la part de l'auteur de créer un contexte ou une atmosphère à son histoire. Il y a donc une relation entre le champ lexical porté par les vocabulaires et cette atmosphère dégagée par le texte.

Dans une approche purement statistique nous avons tenté de retrouver les genres d'animation des films à partir des synopsis et de la présence de termes spécifiques caractéristiques (voir annexe D.2 sur la classification supervisée des synopsis). Dans cette approche un en-

12. Propriété d'un terme qui présente plusieurs sens. Les mots les plus fréquemment utilisés sont le plus souvent polysémiques. En revanche, la monosémie caractérise surtout les vocabulaires scientifiques et techniques.

semble de termes spécifiques sont extraits par une analyse statistique du corpus et de son vocabulaire. Ces termes spécifiques sont consignés dans des références lexicales (liste de mots partageant une même thématique ou de même champ lexical) qui servent ensuite à discriminer les genres des films d'animation. Cependant même si cette approche permet de retrouver des genres comme le genre Publicitaire, Artistique, Expérimental ou Musical, ces lexiques comportent beaucoup de bruits (ici le bruit désigne le fait que les termes appartenant à une référence lexicale ne partagent pas tous le même signifié (ou champ lexical) et/ou que ces signifiés n'ont aucun rapport avec la thématique commune de la référence lexicale). Ce bruit et le manque de sémantique de ces références lexicales contribuent à dégrader les résultats. En effet, on retrouve les termes comme "beau", "espoir" dans la référence lexicale du drame alors que ces termes traduisent plutôt une atmosphère positive. On retrouve également des termes comme "expérience", "fenêtre", "parfois", etc, qui n'ont aucun rapport avec la thématique du drame et qui constituent du bruit sémantique par rapport à la thématique de la référence lexicale. Finalement, l'analyse de ces lexiques (constitués automatiquement par analyse statistique) nous a permis de mettre en évidence des regroupements sémantiques à l'intérieur de ces références lexicales et plus particulièrement dans le cas du lexique associé au genre Dramatique. En effet, des thèmes comme la mort, la guerre, l'horreur, etc, ont pu être retrouvés. Ainsi nous proposons de porter notre analyse sur ce genre particulier du "drame" dont le vocabulaire est facilement reconnaissable car habituellement lié à une atmosphère noire, funèbre et inquiétante. De plus, on « sent bien » que les informations apportées par l'image comme la couleur et ses contrastes peuvent traduire une atmosphère dans la séquence vidéo qui peut être reliée et complétée par une atmosphère traduite par le texte.

Finalement, pour identifier une atmosphère liée au drame à partir du texte nous proposons d'utiliser une analyse thématique des synopsis. Une thématique est une liste de dictionnaires composés de mots ou de regroupements de mots (ou expressions) relevant d'un même thème. L'analyse thématique sert à mesurer la tendance de la thématique à être présente dans le texte. Cette mesure est effectuée par l'intermédiaire de l'intensité lexicale $I_{lex}$ définie comme étant le rapport entre le nombre de mots du texte $Tx$ appartenant à la thématique $Th$ et le nombre total de mots dans ce même texte.

$$I_{lex} = \frac{card(Tx \cap Th)}{card(Tx)} \tag{4.3}$$

avec $I_{lex} \in [0; 1]$. Le calcul de l'intensité lexicale peut être vu comme le calcul du tf dans le calcul du tf-idf.

Malheureusement, une telle ressource linguistique n'a pas été trouvée (WordNet ne dispose pas d'un hyper-concept lié au drame). Nous avons donc créé cette ressource linguistique que nous nommons dictionnaire thématique du drame. Notons que la qualité en terme de précision et d'incertitude de la mesure de l'intensité lexicale est liée à la richesse (en terme de vocabulaire) du dictionnaire thématique mais passe également par la présence de bruit sémantique dans ce dictionnaire (les signifiés de chacun des termes constituant ce dictionnaire doivent appartenir au thème considéré). Pour satisfaire à ces deux contraintes et afin de simplifier la tâche de constitution du dictionnaire par la non prise en compte des différentes flexions des termes, nous utilisons les versions lemmatisées des synopsis. Pour obtenir une richesse suffisante en terme de vocabulaire de notre dictionnaire et obtenir ainsi une plus grande couverture de notre ressource thématique nous utilisons, pour un terme donné, l'ensemble de ses synonymes (termes ayant un signifié (ou sens) identique) obtenus grâce

au dictionnaire des synonymes CRISCO [13]. Ce dictionnaire [Manguin *et al.*, 2004] est mis à disposition du grand public par le Centre de Recherche Inter-langues sur la Signification en COntexte de l'Université de Caen Basse-Normandie. Il contient approximativement 49 000 entrées et 396 000 relations synonymiques issues d'un regroupement de sept dictionnaires classiques : le Bailly, le Benac, le Du Chazaud, le Guizot, le Lafaye, le Larousse et le Robert.

### 4.5.1 Constitution du dictionnaire thématique du drame

Notre objectif est de constituer un dictionnaire thématique semi-automatiquement par le regroupement de synonymes à partir d'un ensemble de termes de départ (notés **germes** par la suite). Pour cela nous avons développé un algorithme de parcours du réseau synonymique du CRISCO qui permet d'extraire l'ensemble des synonymes des germes tout en prenant en compte les points suivants :

- une vérification du parcours a été implémentée pour éviter les rebouclages infinis dans des parties du réseau.

- une profondeur maximale de parcours du réseau a été fixée pour ne pas utiliser de termes dont le signifié serait trop éloigné de la thématique recherchée.

- la vérification de la non extraction d'antonymes. En effet, le parcours d'un tel réseau conduit très rapidement sur des termes antonymes des germes (termes fondateurs de la thématique). Par exemple à partir du germe **drame** nous obtenons les synonymes de premier niveau {*accident, cantate, catastrophe, dramatique, événement, mélodrame, opéra, opéra-comique, oratorio, pièce, pièce de théâtre, théâtre, tragédie, tragi-comédie*} dont les antonymes sont {*comédie, farce, idylle*}. Maintenant si nous parcourons l'ensemble des synonymes du deuxième niveau (les synonymes des synonymes du germe) nous obtenons pour le terme **théâtre** la liste de ses synonymes {*amphithéâtre, arène, boui-boui, boulevard, bunraku, café-concert, café-théâtre,* ***comédie****, compagnie, décor, drame, emplacement, endroit, farce, kabuki, lieu, littérature, mimesis, miracle, mystère, nô, oeuvre, opéra, opéra-comique, pigeonnier, planches, plateau, salle, scène, site, studio, tréteaux*}. On remarque que parmi les synonymes du deuxième niveau on obtient le terme **comédie** qui est un antonyme du germe (**drame**). Une telle vérification est donc nécessaire pour ne pas regrouper les germes et leurs antonymes.

Pour lancer le processus de constitution du dictionnaire il faut au préalable définir un ensemble de germes. Pour cela nous cherchons parmi l'ensemble des 5804 synopsis de la base des films inscrits au festival (voir annexe B) le vocabulaire spécifique des synopsis des films dont le genre déclaré est le drame. Pour obtenir ce vocabulaire nous cherchons à obtenir la liste des mots qui pour une catégorie ou un contexte $\lambda$ (en l'occurrence le drame déclaré) semblent être sur-représentés par rapport aux autres catégories ou contextes. Cette liste de mots est construite à partir de l'indice de spécificité de chaque modalité. Cet indicateur est le rapport entre le nombre d'utilisations observées et le nombre théorique d'utilisations tel qu'il résulterait d'un emploi proportionnel au nombre total de mots prononcés par la catégorie considérée. Les calculs de cet indice de spécificité s'apparentent aux calculs effectués pour le test du $Chi^2$ [Sphinx, 2009]. Il s'agit de mettre en évidence des écarts à une répartition

13. http://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

de référence. On procède en calculant un effectif théorique répondant à une hypothèse de répartition proportionnelle des éléments étudiés. L'écart à la référence est mis en évidence par le rapport entre l'effectif théorique et celui que l'on observe. Si on note $N$ le nombre total de mots dans le corpus, $m$ le mot utilisé, $\lambda$ la catégorie considérée, $N_m$ le nombre de fois où le mot $m$ est utilisé par toutes les catégories confondues, $N_\lambda$ le nombre total de mots dans la catégorie $\lambda$ et $N_m(\lambda)$ le nombre de fois où le mot $m$ est utilisé dans la catégorie $\lambda$ alors le nombre théorique d'utilisations tel qu'il résulterait d'un emploi proportionnel au nombre total de mots prononcés par la catégorie considérée est égal à :

$$N_{Th}(m) = N_\lambda * \frac{N_m}{N} \tag{4.4}$$

et donc l'indice de spécificité est égal à :

$$I^\lambda_{spe}(m) = \frac{N_m(\lambda)}{N_{Th}(m)} \tag{4.5}$$

- si les 2 effectifs sont identiques, le rapport est égal à 1, la répartition est proportionnelle.
- si l'effectif réel est supérieur à l'effectif théorique, l'élément considéré est sur-représenté dans la catégorie considérée et le rapport est supérieur à 1.
- si l'effectif réel est inférieur à l'effectif théorique, l'élément considéré est sous-représenté dans la catégorie considérée et le rapport est inférieur à 1.
- si un mot est spécifique à une seule des catégories alors son indice de spécificité tend vers l'infini. Avec cette mesure on peut par exemple extraire les mots exclusifs à une catégorie.

Nous appliquons ce calcul sur les synopsis lemmatisés et débarrassés des mots outils et nous conservons les mots dont l'indice de spécificité $I^\lambda_{spe}$ est supérieur à 1.5. Nous obtenons une liste de plus de 1900 termes dont 370 sont exclusifs ($I^\lambda_{spe} = \infty$) à cette catégorie du drame. C'est par exemple les termes comme *affreusement, angoisser, barjot, conspiration, geisha, laideur, poignarder, vulnérabilité*. Parmi ces termes beaucoup traduisent une atmosphère noire, dramatique, morbide ou violente (voir annexe D.4) même si il y a beaucoup de bruit (termes ne partageant pas le champ lexical du drame). Après filtrage manuel [14] de cette liste de mots spécifiques nous obtenons une liste de plus de 80 germes comme : *drame, mort, macabre, accident, détruire, terrible, violent, violer, sinistre, sacrifier, etc.* (voir annexe D.5). Le dictionnaire est constitué de l'ensemble des synonymes de niveau 1 et 2 de ces 80 germes puis il est filtré manuellement afin de supprimer les termes aberrants (bruit). Finalement ce dictionnaire de la thématique du drame est constitué d'un peu moins de 800 termes.

### 4.5.2 Test et résultats

Pour vérifier le pouvoir discriminant de cette mesure nous calculons pour chaque synopsis l'intensité thématique du drame (intensité lexicale associée au thème du drame), puis nous calculons la moyenne des intensités thématiques pour chacune des catégories du genre déclaré sans tenir compte des non-réponses. On voit sur la figure 4.24 que cette intensité thématique

14. nous ne conservons que les termes qui se rapportent au drame c'est-à-dire à des concepts comme la mort, la guerre, la souffrance, la tristesse, l'horreur, la noirceur, etc.

est importante et significative (les noms des critères discriminants sont encadrés et correspondent à des moyennes significativement différentes de l'ensemble de l'échantillon au risque de 95% (test de student)) dans le cas des genres **Humour noir**, **Policier**, **Drame** et **Satire**.

| Genre | I Drame |
|---|---|
| __litterature | 2.24 |
| _argent | 0.00 |
| _educatif | 4.56 |
| _provocateur | 5.25 |
| Aventure | 4.30 |
| Comédie | 5.93 |
| Comédie_musicale | 2.24 |
| Conte | 3.92 |
| Documentaire | 3.09 |
| Drame | 8.28 |
| Epopée | 4.19 |
| Erotique | 3.44 |
| Expérimental | 3.95 |
| Fantastique | 5.26 |
| Humour | 4.43 |
| Humour_noir | 9.52 |
| Musical | 4.10 |
| Policier | 9.16 |
| Promotion | 0.00 |
| Propagande | 4.66 |
| Publicité | 1.22 |
| Satire | 6.16 |
| Science-fiction | 6.03 |
| TOTAL | 4.78 |

FIGURE 4.24 – Moyenne des intensités thématiques du drame en fonction du genre déclaré

Cependant même si le genre humour noir a un indice thématique important cette catégorie de film est très minoritaire puisque seulement 2 films sur 5804 ont été déclaré de ce genre. Pour évaluer le pouvoir discriminant de cette mesure thématique, nous décidons de classifier naïvement et simplement les synopsis suivant cette intensité thématique avec la règle suivante :

$$\textbf{\textit{SI}}\ 8.3 \leq I_{drame}(S)\ \textbf{\textit{ALORS}}\ Genre_{predit}(S) \leftarrow DRAME.$$

Si l'intensité thématique dramatique $I_{drame}$ du synopsis $S$ est supérieure au seuil de 8.3% (moyenne des intensités thématique associées au Drame) alors on considère que le genre du film d'animation est le *drame*. Nous comparons les résultats de cette règle avec le genre déclaré. Si le genre déclaré est le drame alors la classifieur a retrouvé le genre du film sinon il s'est trompé. Nous obtenons la matrice de confusion (voir tableau 4.6) où chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe déclarée (ou de référence).

Nous pouvons voir sur la figure 4.25 et avec la matrice de confusion 4.6 que le nombre de **Faux Positifs** (FP) et de **Faux Négatifs** (FN) qui représentent l'erreur globale de la prédiction sont relativement importants (24%). A partir de cette matrice de confusion nous calculons les deux indicateurs que sont la précision et le rappel :

| Déclaré \ Estimé | *NonDrame* | *Drame* |
|---|---|---|
| *NonDrame* | 2892 *(VN)* | 614 *(FP)* |
| *Drame* | 363 *(FN)* | 194 *(VP)* |

TABLE 4.6 – Matrice de confusion sur la prédiction du Drame. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

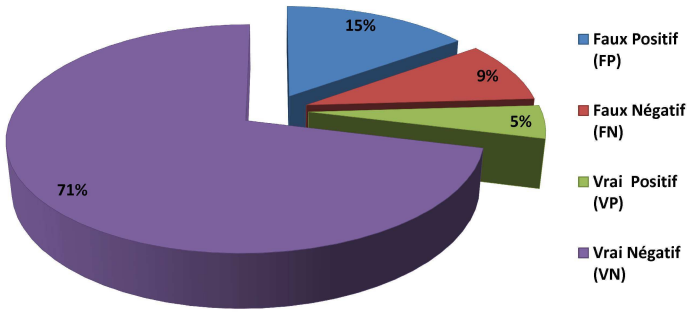



FIGURE 4.25 – Représentation graphique des résultats sur la prédiction du Drame

$$\text{Précision} = \frac{VP}{VP + FP} = \frac{194}{194 + 614} = 24\%$$
$$\text{Rappel} = \frac{VP}{VP + FN} = \frac{194}{194 + 363} = 34\%$$

Nous utilisons également le F-score (ou F-mesure) mesure qui combine la précision et le rappel :

$$F_{score} = 2 * \frac{P * R}{P + R} = 2 * \frac{24 * 34}{24 + 34} = 28\%$$

On remarque que ces taux sont relativement faibles et que les résultats ne sont pas bons (loin de 100%). Cependant ils sont nettement améliorés (augmentation de 385% du rappel et de 71% de la précision) par rapport à l'approche purement statistique (voir figure D.3). Les causes de ces faibles résultats sont les suivantes :

– Ces faibles résultats sont liés à l'utilisation de vocabulaires à connotation dramatique dans un contexte qui ne l'est pas ou bien à des atmosphères dramatiques qui ne sont perceptibles qu'à travers la compréhension de l'énoncé et non par l'utilisation d'un vocabulaire spécifique. C'est par exemple le cas dans les synopsis des films suivants :

**Le bonhomme de neige (1960)** : *Le printemps annonce la fin de l'idylle entre une fillette et un bonhomme de neige. Lorsque ce dernier a tout à fait fondu, demeure*

*à sa place une fleur, hommage d'affection.*

**Le manège (1979)** : *Dans une ville, par une nuit de violent orage, titubant sous de grands parapluies noirs, des silhouettes convergent vers un mystérieux manège de chevaux de bois.*

**Signal (1983)** : *Un train s'apprête à partir. On entend le signal de départ et les voyageurs se précipitent vers les wagons, mais un autre signal les ramène vers le quai. Les signaux se succèdent, créant un énorme affolement.*

– Ces faibles résultats sont liés également à l'utilisation d'une vérité terrain de mauvaise qualité. En effet le champ des genres déclarés de la base de synopsis est pris comme vérité terrain ; or ce champ est souvent mal renseigné ce qui conduit à augmenter le nombre de Faux Positifs. Par exemple lorsque l'on s'intéresse à ces derniers (voir figure 4.26) on remarque que beaucoup de ces synopsis sont déclarés comme **Policier**, **Satire**, **Humour** alors qu'ils dégagent, par leur lexique, une atmosphère dramatique. Dans ce dernier cas cette dualité entre humour et drame peut être voulue par l'auteur mais correspondrait probablement plus au genre **Humour noir**.

**The Old Man and the Flower (1960)** : Genre déclaré : **Aventure**
*L'histoire d'un vieux bonhomme seul. Une fleur le remarque et lui donne de la joie mais il la perd.*

**La pie voleuse (1964)** : Genre déclaré : **Comédie Musicale**
*La guerre terminée, ne trouvant chez lui que ruines, un soldat descend aux enfers. Le diable l'engage pour entretenir le feu sous les chaudrons. De ceux-ci sortent successivement un capitaine, un maréchal et le roi, qui supplient le soldat de les libérer. Il refuse, les enferme soigneusement dans leur chaudière et décide de remonter sur Terre où, sans ces tyrans, la vie est maintenant belle.*

**La métamorphose de M. Samsa (1977)** : Genre déclaré : **Fantastique**
*M. Samsa, en s'éveillant un matin, découvre qu'il est transformé en cafard. Il connaît la souffrance de la réclusion et celle d'être rejeté de ceux dont il a jusqu'alors partagé l'existence..*

**Les jeux des anges (1983)** : Genre déclaré : **Satire**
*Une vision où la tragédie et l'horreur de notre monde moderne, dans ce qu'il engendre de violence et de sévices, sont symbolisées et suggérées beaucoup plus que décrites : un reportage dans la cité des anges.*

**Taxi de nuit (1996)** : Genre déclaré : **Policier**
*Parce qu'il s'est aventuré un soir dans la 42e rue et qu'il a rencontré Slacks, un chauffeur de taxi va être condamné à mort. C'est de prison qu'il nous raconte son histoire.*

| TEST_DRAME<br>Genre | VN | FP | FN | VP | TOTAL |
|---|---|---|---|---|---|
| _argent | 100% ( 2) | 0.0% ( 0) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 2) |
| __litterature | 100% ( 6) | 0.0% ( 0) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 6) |
| Promotion | 100% ( 8) | 0.0% ( 0) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 8) |
| Drame | 0.0% ( 0) | 0.0% ( 0) | 65.2% ( 363) | 34.8% ( 194) | 100% ( 557) |
| Humour_noir | 50.0% ( 1) | 50.0% ( 1) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 2) |
| Comédie | 66.7% ( 8) | 16.7% ( 2) | 8.3% ( 1) | 8.3% ( 1) | 100% ( 12) |
| Science-fiction | 77.3% ( 17) | 13.6% ( 3) | 4.6% ( 1) | 4.6% ( 1) | 100% ( 22) |
| Publicité | 95.8% ( 68) | 4.2% ( 3) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 71) |
| Epopée | 73.3% ( 11) | 26.7% ( 4) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 15) |
| Erotique | 78.9% ( 56) | 11.3% ( 8) | 7.0% ( 5) | 2.8% ( 2) | 100% ( 71) |
| Conte | 84.4% ( 54) | 14.1% ( 9) | 0.0% ( 0) | 1.6% ( 1) | 100% ( 64) |
| Propagande | 71.4% ( 45) | 15.9% ( 10) | 7.9% ( 5) | 4.8% ( 3) | 100% ( 63) |
| Policier | 42.9% ( 12) | 39.3% ( 11) | 10.7% ( 3) | 7.1% ( 2) | 100% ( 28) |
| Documentaire | 82.3% ( 107) | 12.3% ( 16) | 4.6% ( 6) | 0.8% ( 1) | 100% ( 130) |
| Musical | 72.2% ( 96) | 12.8% ( 17) | 9.8% ( 13) | 5.3% ( 7) | 100% ( 133) |
| Comédie_musicale | 90.7% ( 176) | 9.3% ( 18) | 0.0% ( 0) | 0.0% ( 0) | 100% ( 194) |
| _provocateur | 68.9% ( 71) | 25.2% ( 26) | 4.9% ( 5) | 1.0% ( 1) | 100% ( 103) |
| _educatif | 78.1% ( 243) | 19.6% ( 61) | 1.9% ( 6) | 0.3% ( 1) | 100% ( 311) |
| Fantastique | 72.1% ( 269) | 19.0% ( 71) | 6.7% ( 25) | 2.1% ( 8) | 100% ( 373) |
| Satire | 71.3% ( 219) | 24.4% ( 75) | 2.6% ( 8) | 1.6% ( 5) | 100% ( 307) |
| Aventure | 78.3% ( 376) | 16.7% ( 80) | 4.2% ( 20) | 0.8% ( 4) | 100% ( 480) |
| Expérimental | 76.0% ( 681) | 14.4% ( 129) | 7.7% ( 69) | 1.9% ( 17) | 100% ( 896) |
| Humour | 77.4% ( 996) | 17.3% ( 222) | 3.4% ( 44) | 1.9% ( 25) | 100% (1287) |
| TOTAL | 71.2% (3522) | 15.1% ( 766) | 8.9% ( 574) | 4.8% ( 273) | 100% (5135) |

FIGURE 4.26 – Répartition des résultats de la prédiction du Drame suivant le genre déclaré. *Les effectifs sont entre parenthèses. Les effectifs sont supérieurs aux nombres d'observations en raison de réponses multiples (plusieurs genres par synopsis)*

### 4.5.3 Conclusion partielle

Finalement, l'utilisation du lexique seul n'est pas suffisante pour retrouver à coup sûr les genres d'animation ; il est nécessaire de prendre en compte d'autres informations. L'amélioration des résultats précédents passe par exemple par la diminution des fausses détections (FP) qui peut être réalisée grâce à :

- La détection d'autres atmosphères comme le Policier, l'Humour (voir les annexes D.3.2 et D.3.3 pour l'analyse thématique de ces atmosphères) qui sont à l'origine des fausses détection permettrait de ne pas classer ces synopsis comme Dramatique.

- L'apport d'informations non textuelles issues du film lui même comme les informations de couleur ou d'activité permettrait de compléter les informations textuelles.

## 4.6 Conclusion

Dans ce chapitre nous avons vu comment extraire de l'information à partir de textes pour caractériser les films d'animation. Cette caractérisation des films à travers les synopsis s'opère suivant deux niveaux. En effet, nous avons vu comment extraire une information plutôt locale (c'est-à-dire localisée dans une sous-séquence du film) par l'intermédiaire du scénario actanciel. Ce dernier décrit l'action, les protagonistes et le contexte de l'histoire. Nous avons vu également, comment extraire une information globale au travers de l'atmosphère dégagée par le film. Cette information (qui est liée au genre d'animation) est étudiée dans nos travaux

au travers de la thématique du drame. Cependant les résultats de la recherche d'atmosphère dramatique au travers de l'analyse des lexiques ne sont pas satisfaisants et l'apport d'autres sources d'information est nécessaire pour pouvoir les améliorer. Cette utilisation conjointe de différentes sources d'information à des fins de caractérisation des films est abordée dans la partie concernant la fusion d'information dans le chapitre suivant.

# Troisième partie

# Fusion d'information

Chapitre 5

# La fusion d'information entre le texte et l'image

**Résumé :** Dans ce chapitre sont abordées les problématiques de fusion d'information entre les deux sources d'informations hétérogènes que sont les images et le texte. En effet, notre approche permet d'utiliser conjointement les informations issues de l'analyse des séquences d'images et de l'analyse des synopsis dans le but de caractériser les films d'animation. Cette caractérisation est présentée suivant deux niveaux différents à travers la caractérisation globale et locale de la séquence vidéo. La caractérisation globale du film est basée sur l'analyse de l'atmosphère dégagée à partir des images et des textes. La fusion de ces informations est obtenue en s'appuyant sur une expertise du domaine et implémentée par des systèmes de fusion floue. La caractérisation locale du film quant à elle est basée sur l'analyse de l'activité dans les images et sur la description textuelle de cette activité à partir du scénario actanciel.

Nous avons vu dans les précédents chapitres les méthodes qui permettent l'extraction de caractéristiques à partir des films d'animation. Ces informations issues de sources différentes vont dans ce chapitre être utilisées conjointement afin de caractériser le film de façon plus pertinente que si l'on ne disposait que d'une source d'information. Ainsi, nous allons voir dans ce chapitre les caractéristiques d'un système de fusion d'information et sa mise en œuvre dans notre contexte applicatif.

## 5.1 État de l'art sur la fusion d'information

Qu'est ce que la fusion d'information ? Pour répondre à cette question prenons comme exemple le plus complexe et le plus sophistiqué des systèmes de fusion d'information, c'est-à-dire le cerveau. Le cerveau réalise un processus cognitif extraordinaire lui permettant d'ana-

lyser et d'agréger des centaines d'informations de nature différentes arrivant en parallèle et à chaque instant. La combinaison de ces informations issues de nos sens, nous permet de réaliser une interprétation de l'environnement et de prendre des décisions en conséquence. Par exemple, si nous nous intéressons à la fonction de stabilisation (c'est-à-dire rester debout) grossièrement trois informations sont utilisées : la position transmise par l'oreille interne (sorte de niveau à bulle), la position dans l'espace transmise par la vue et enfin des informations sur l'environnement sur lequel nous évoluons transmises par l'intermédiaire de certains capteurs du pied. C'est la combinaison de ces trois informations qui nous permet de vaincre la gravité et de nous maintenir en équilibre. L'utilisation de ces différentes sources d'information est nécessaire pour garantir au processus de maintien en équilibre une performance optimale quelque soit la situation dans laquelle se trouve le corps et qui peut être affectée par différents facteurs externes ou internes. En effet, les conditions extérieures comme un sol glissant, la présence de brouillard ou d'obscurité peuvent dégrader les informations utilisées par le cerveau et conduisent à une incertitude sur l'environnement dans lequel nous évoluons. L'évaluation de la situation peut également être affectée par des conditions internes au système de fusion c'est par exemple l'entraînement d'un funambule ou d'un patineur mais cela peut également être des dégradations du système lui même par exemple au manque de sobriété de la personne.

Finalement la qualité du résultat de la fusion de ces informations a des conséquences importantes sur les décisions et les actions à entreprendre. La prise en compte de ces imperfections (imprécision, incertitude, etc.) par le système de fusion augmente sa robustesse et permet la prise de décision dans un contexte perturbé. Ainsi, toutes ces prouesses réalisées par notre cerveau n'ont cessé d'inspirer les techniques développées pour la fusion d'information.

D'un point de vue technique, depuis de nombreuses années, les données issues de sources diverses et variées que l'on nomme par exemple capteur dans les domaines industriel ou médical ne cessent d'augmenter. Ces données, bien que porteuses d'une information, peuvent être dégradées et manquer de précision, de fiabilité ou d'interprétabilité. Se pose alors le problème de la prise en compte de l'ensemble de ces données pour tendre vers l'objectif voulu (réduire l'incertitude sur l'information résultante en est un exemple).
Le domaine de recherche de la fusion de données voit le jour dans les années 60 pour répondre au besoin de combiner ces données afin d'améliorer la prise de décision. Les méthodes mises en œuvre sont diverses et sont généralement basées sur des approches probabilistes [Goodman *et al.*, 1997].

L'utilisation de la micro informatique et l'évolution des technologies ont conduit à l'utilisation de capteurs de plus en plus « intelligents », obtenant ainsi des données plus élaborées dont le niveau sémantique est parfois plus élevé que les données brutes. L'utilisation de la sémantique permettant une meilleure interprétabilité des données conduit le domaine de la fusion de données à muter et à devenir le domaine de la fusion d'information dans les années 90. Aujourd'hui un grand nombre d'applications dans des domaines différents emploient la fusion d'information pour répondre à des problématiques de plus en plus complexes. De plus, l'utilisateur devient un élément clef dans la chaîne de fusion puisque l'information issue du système de fusion doit être interprétable, si possible avec un niveau d'abstraction plus élevé que les informations traitées. Ces systèmes coopératifs de fusion d'information permettent ainsi à l'utilisateur d'interagir avec le système de fusion assurant une coopération

entre l'Homme et la machine.

Dans la littérature, différentes définitions de la fusion d'information ont été proposées et mettent en avant des aspects différents comme [Valet, 2001] :

- L'amélioration de la qualité des informations. Le résultat de la fusion permet d'obtenir une meilleure information (en termes de confiance, de certitude ou de robustesse).
- L'obtention d'une nouvelle information. Le résultat de la fusion doit apporter une information qu'il n'est pas possible d'obtenir à partir des données étudiées séparément.
- L'élévation du niveau sémantique du résultat obtenu.
- La prise de décision à partir des informations fusionnées.

D'autres définitions sont plus globales, comme celle proposée par Lucien Wald [Wald, 1999] :

**"Data fusion is a formal framework in which are expressed the means and tools for the alliance of data originating from different sources. Data fusion aims at obtaining information of greater quality ; the exact definition of "greater quality" will depend upon the application."**

*« La fusion de données constitue un cadre formel dans lequel s'expriment les moyens et techniques permettant l'alliance des données provenant de sources diverses. Le but de la fusion de données est d'obtenir une information de meilleure qualité ; la définition exacte de « meilleure qualité » dépendra de l'application. »*

Cette définition bien que générale est intéressante car elle fait bien apparaître les points importants d'un système de fusion d'information comme la définition d'un **cadre formel** pour combiner des sources d'informations variées.

Historiquement, la fusion d'information a été initialement réservée au domaine militaire pour des tâches de détection [Bastière, 1998, Li *et al.*, 2002, Maussang *et al.*, 2008], d'identification et de suivi [Wu et Zhu, 1999, Volgyesi *et al.*, 2007] de cibles puis elle s'est très vite étendue à d'autres domaines comme le domaine de l'aéronautique et du spatial [Volponi *et al.*, 2003] avec par exemple l'imagerie satellitaire [Bujor *et al.*, 2002] ou la commande d'engins (robots spatiaux, pilotage automatique d'avion [Korn, 2006], etc.). Elle s'est également développée dans le domaine médical [Abbod *et al.*, 2001] avec l'imagerie médicale [Barra, 2000] et la détection de pathologies ou avec l'assistance robotisée dans les opérations chirurgicales [Troccaz, 2006, Cinquin et Troccaz, 2009]. Elle envahit aussi le domaine de l'assistance à l'être humain avec par exemple l'aide à la conduite ou l'aide au contrôle comme l'aiguillage du ciel ou la gestion de l'énergie électrique [Besada *et al.*, 2004]. Les domaines concernés par cette thématique de recherche sont nombreux et ne cessent de s'accroitre avec l'augmentation du nombre de sources d'informations (issues de capteurs intelligents, d'images sophistiquées, etc.), du développement d'algorithmes et de l'accroissement des capacités de calcul. Cet engouement pour la fusion d'information est soutenu par le besoin d'information

de haut niveau sémantique et de l'intégration de l'homme dans les systèmes.

### 5.1.1 Les objectifs d'un système de fusion

D'après la définition de **Wald** un système de fusion d'information permet d'obtenir une information dont la qualité dépend de l'application. Cela signifie que le développement du système de fusion est lié aux objectifs recherchés en sortie du système. Les principaux objectifs sont décrits dans [Valet, 2001, Dubois et Prade, 2004] et peuvent être les suivants :

**Réduction de la dimensionnalité et augmentation du niveau d'abstraction :** Cela consiste à synthétiser dans l'information de sortie les informations en entrée. L'utilisateur dispose alors d'une information synthétique plus facile à interpréter. Deux aspects sont à considérer : le premier consiste en la réduction des dimensions de l'espace de représentation de l'information de sortie par rapport à celui de l'espace d'entrée. Le deuxième aspect est que la fusion va permettre d'augmenter le niveau d'abstraction qui caractérise l'information. Par exemple dans le cas d'une prise de décision à partir de plusieurs informations, la décision a le degré d'abstraction maximal alors que les données issues directement des capteurs ont le plus faible degré d'abstraction. Cet objectif apparaît dans la littérature consacrée à la fusion d'information comme le plus repandu [Valet *et al.*, 2001].

**Amélioration de la précision et de la certitude** de l'information : La définition de **Wald** insiste sur la nécessité d'obtenir en sortie une information de meilleure qualité qu'en entrée. Cette notion de meilleure qualité est liée aux défauts intrinsèques des capteurs et des algorithmes de traitement des informations. En effet, une imprécision, une incertitude, un retard ou un manque de données apportent une imperfection des informations qu'il est nécessaire de réduire. La réduction de l'imprécision et/ou de l'incertitude de l'information est un aspect de la qualité de l'information fusionnée. Ces notions de précision et de certitude sont fortement liées : plus une information est précise et plus elle risque d'être incertaine et vice-versa. L'imprécision est liée à l'estimation de la différence entre la mesure $d$ provenant du capteur et la valeur réelle inconnue $x$ à mesurer. L'incertitude est un doute sur la réalité des hypothèses ou la véracité d'une mesure (par exemple on doute des résultats d'un sondage réalisé sur seulement 100 personnes lors d'une élection présidentielle).

**Robustesse** de l'information : La robustesse du résultat de fusion est liée à la fiabilité de l'information obtenue en sortie du système de fusion lorsque les informations d'entrée sont bruitées. Ces informations d'entrée peuvent subir d'éventuelles détériorations ou même être absentes. Pour tester la robustesse des systèmes de fusion une méthode consiste à ajouter du bruit aux données d'entrée puis à comparer les indicateurs de performance de l'information fusionnée à ceux obtenus sans bruit.

### 5.1.2 Structure d'un système de fusion

Habituellement la fusion d'information est présentée comme un système composé de quatre étapes [Rombaut, 2001, Valet, 2001, Jullien, 2008]. Ces étapes sont représentées chronologiquement sur la figure 5.1, l'information circulant de gauche à droite. La première étape nommée **acquisition** de l'information recueille les caractéristiques du système physique à étudier. Ensuite ces informations sont mises en forme pour la fusion dans l'étape de **représentation**. Une fois les informations mises en forme elles sont agrégées par l'utilisation d'opérateurs de fusion dans l'étape de **combinaison**. Finalement une étape d'**interprétation** permet d'adapter l'information fusionnée à son usage ultérieur.

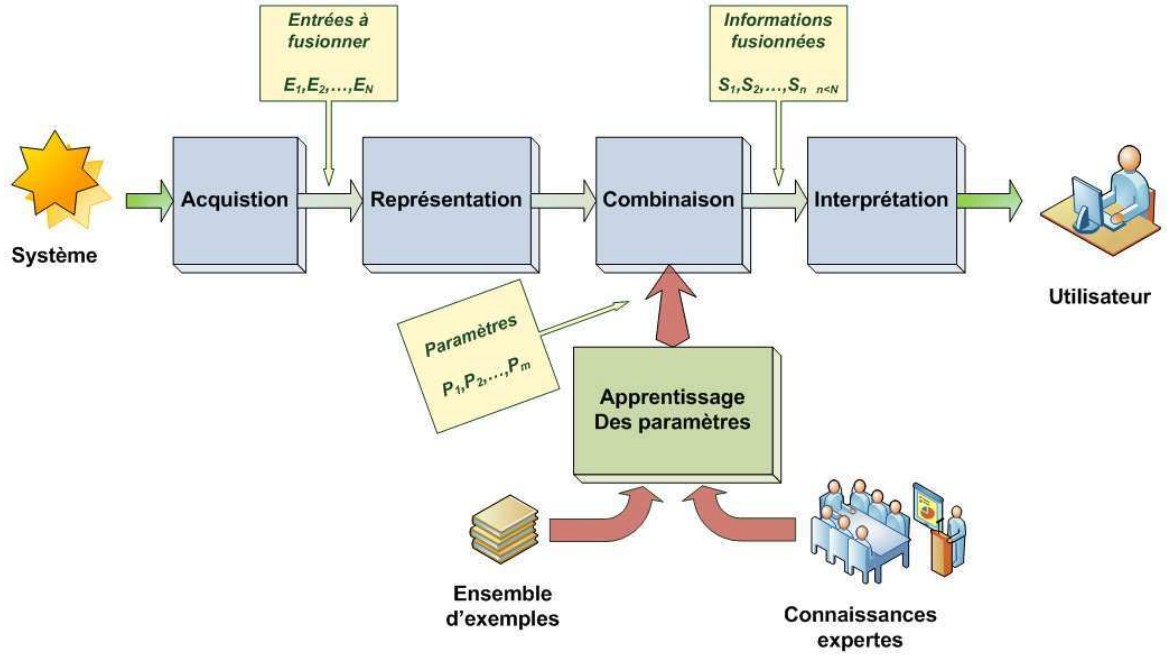



FIGURE 5.1 – Structure d'un système de fusion d'information

### 5.1.3 L'acquisition de l'information

La première étape de la figure 5.1 est l'**acquisition** de l'information. Elle a pour but de recueillir des informations sur le système que l'on souhaite étudier. Le rôle du concepteur est important dans cette étape. En effet c'est lui qui décide quelles sont et en quelle quantité les informations physiques pertinentes nécessaires pour obtenir l'information de sortie. On parle de niveau d'abstraction pour caractériser cette information [Rombaut, 2001]. Le niveau le plus faible correspond au signal issu du capteur ou donnée brute. Le second niveau est le niveau de l'attribut, il correspond aux résultats des traitements sur le signal (par exemple l'extraction des régions ou des contours en traitement de l'image). Le niveau objet permet, à partir du regroupement des attributs, de caractériser les objets physiques qui sont observés. Finalement le dernier niveau est le niveau de la décision : il s'agit du niveau sémantique le plus élevé, où l'on cherche à identifier, classifier, reconnaître ce qui est observé.

Ces informations d'entrée ne sont pas forcément isolées ou indépendantes. En effet il peut

exister un lien ou une dépendance entre ces informations [Valet, 2001]. Par exemple, la **complémentarité** entre deux sources d'information est employée lorsque ces sources mesurent des informations différentes (de par leur nature ou bien leur plage de variation) concernant un même phénomène. La **redondance** entre données exprime quant à elle le caractère identique des informations sur le même phénomène ou le même objet. Cette propriété est généralement exploitée pour améliorer la qualité des informations en termes de précision et d'incertitude. Le terme de recouvrement est aussi utilisé lorsque les informations se recoupent partiellement. Les informations sont dites en **concordance** quand rien n'empêche qu'elles soient vraies simultanément. Au contraire elles sont dites en **conflit** lorsque leurs affirmations ne sont pas compatibles (elles ne peuvent être vraies simultanément). On retrouve aussi la notion de **coopérativité** entre informations qui consiste à faire coopérer différentes informations pour atteindre l'objectif de fusion.

De nombreuses méthodes tentent de mesurer cette dépendance entre les informations. Dans [Jullien, 2008] l'auteur distingue deux types de dépendance, la dépendance fonctionnelle qui correspond au cas où les relations peuvent être observées de manière déterministe et la dépendance statistique portant sur les relations entre les distributions statistiques des données. Dans ce dernier cas de nombreuses méthodes existent, souvent basées sur la quantification d'une relation statistique.

**Le coefficient de corrélation linéaire de Bravais-Pearson** est la méthode la plus connue. Elle revient à caractériser la relation affine entre la variable $X(x_1, \ldots, x_n)$ et la variable $Y(y_1, \ldots, y_n)$ (voir equation 5.1)

$$r_p = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{5.1}$$

$r_p$ est égal à 1 dans le cas où l'une des variables est une fonction affine croissante de l'autre variable (ou égale à -1 dans le cas où la fonction affine est décroissante). Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte. Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes. Cependant la réciproque n'est pas vraie, car le coefficient de corrélation indique uniquement une dépendance linéaire. La corrélation de Spearman est utilisée lorsque deux variables statistiques semblent corrélées sans que la relation entre les deux variables soit de type affine. Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables, mais entre les rangs de ces valeurs.

**L'information mutuelle** de deux variables aléatoires est une quantité mesurant la dépendance statistique de ces variables. L'information mutuelle d'un couple (X,Y) de sources représente le degré d'interaction entre ces deux sources d'information. Elles sont dites indépendantes ($I(X,Y) = 0$) si la réalisation de l'une n'apporte aucune information sur la réalisation de l'autre.

$$I(X,Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)\,P(y)} \tag{5.2}$$

**Les Q statistiques** permettent de mesurer les dépendances entre les résultats de deux classifieurs. Dans [Kuncheva et Whitaker, 2003] l'auteur présente 10 mesures de diversité entre classifieur. La mesure notée $Q_{i,k}$ permet de quantifier la diversité des informations apportées par deux classifieurs ($D_i$ et $D_k$) à partir des résultats de classification (voir équation 5.3 et figure 5.2). La mesure $Q_{i,k}$ varie de -1 à +1. Lorsque $Q_{i,k}$ vaut 0 les classifieurs sont indépendants. Lorsque $Q_{i,k}$ tend vers +1 les classifieurs ont tendance à reconnaitre les même objets. Lorsque $Q_{i,k}$ tend vers -1 les classifieurs ont tendance à reconnaitre des objets différents. En notant $N^{11}$ le nombre d'éléments correctement classifiés par les deux classifieurs, $N^{00}$ le nombre d'éléments mal classifiés par les deux classifieurs, $N^{01}$ et $N^{10}$ le nombre d'éléments correctement classifiés par seulement l'un des deux classifieurs, on a alors :

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{5.3}$$

Cette mesure sera utilisée par la suite.

| | $D_k$ correct (1) | $D_k$ wrong (0) |
|---|---|---|
| $D_i$ correct (1) | $N^{11}$ | $N^{10}$ |
| $D_i$ wrong (0) | $N^{01}$ | $N^{00}$ |

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

FIGURE 5.2 – Résultats de la classification

### 5.1.4 La représentation de l'information

Les informations acquises ne sont pas toujours homogènes et sont définies dans leur espace propre (celui du capteur). Cependant, la fusion de ces informations ne peut se faire que dans un espace de représentation commun $\Omega$. Cette étape permet de conditionner les informations issues de l'acquisition pour obtenir une représentation de l'information la plus proche possible de la réalité tout en ayant une complexité réduite au niveau du support de représentation $\Omega$. Généralement, les supports mathématiques les plus utilisés sont la théorie des probabilités, la théorie des sous-ensembles flous et la théorie de l'évidence [Valet, 2001]. Le format de cette information après conversion (les méthodes de conversion sont obtenues par **apprentissage** ou par **expertise**) peut être une valeur ou une distribution numérique ou symbolique. Notons aussi que cette étape de représentation est parfois appelée *alignement des sources* (dans [McDaniel, 2001], nous avons la définition suivante : "Data Alignment : Normalization of data with respect to time, space, and units to permit common data processing").

### 5.1.5 La combinaison de l'information

Une fois que les informations sont représentées dans un même espace commun normalisé dans l'espace, le temps et les unités, l'opération d'agrégation peut être appliquée. Ses

opérateurs de combinaison sont nombreux et plus ou moins complexes : ils peuvent être une simple moyenne, l'utilisation de règles floues, de classifieur type SVM ou réseaux de neurones. Un exemple d'opérateur assez simple et très souvent utilisé dans la vie de tous les jours est la moyenne pondérée. Prenons par exemple un conseil de classe qui se réunit pour statuer sur le passage en classe supérieure des élèves d'une classe (objectif de la fusion). L'élève constitue le "système" (pour la figure 5.1) sur lequel un certain nombre de mesures sont effectuées durant l'année (devoirs notés) sur des propriétés différentes (matières enseignées) de ce système. Le conseil désire obtenir en sortie du système de fusion un indicateur global permettant de prendre une décision quant au passage de l'élève dans la classe supérieure. Finalement, l'objectif de ce système de fusion d'information consiste en la réduction de la dimensionnalité à partir d'informations d'entrée complémentaires. L'opérateur généralement choisi est la moyenne pondérée qui prend en entrée les différentes mesures exprimées dans un même espace commun (notes de 0 à 20) et qui, par une opération arithmétique, permet de synthétiser les différentes informations. Évidement les performances de ce système dépendent fortement du choix des poids associés à chacune des sources d'information (le coefficient des matières). Dans cette étape de combinaison le concepteur du système de fusion doit apporter une attention particulière au réglage des paramètres. Deux approches sont possibles (voir figure 5.1) :

- L'utilisation d'une expertise et des connaissances sur le système permet de fixer *a priori* les paramètres de l'étape de combinaison. C'est la méthode généralement admise dans notre exemple où des experts de l'enseignement fixent les coefficients des matières enseignées (coefficient de 7 pour les mathématiques au bac Scientifique par exemple).

- L'utilisation de méthodes automatiques (ou semi automatiques) permet, à partir d'un ensemble d'exemples, d'apprendre les paramètres de l'étape de combinaison. Cette approche est utilisée lorsque les paramètres de fusion ne sont pas connus *a priori*. Cette approche n'est généralement pas utilisée dans l'exemple précédent.

Le choix des paramètres est en réalité influencé par les dépendances qui existent entre l'information en sortie et les informations en entrée du système de fusion (voir figure 5.3). Ces dépendances *fonctionnelles* peuvent modéliser différents liens sémantiques entre ces informations [Jullien, 2008] :

- Des liens de nature **physique**, issus des lois de la physique qui gouvernent le phénomène étudié et qui relient les informations d'entrée à celle de sortie.

- Des liens de nature de **confiance**. L'information de sortie apporte une plus grande confiance à certaines sources d'entrée. Ainsi des liens de dépendance apparaissent. Par exemple pour réaliser une synthèse des sondages réalisés durant une élection présidentielle on accordera plus de confiance à une étude faite par une société reconnue et neutre (type Ipsos) qu'à un sondage indépendant d'un parti politique par exemple.

- Des liens de nature **préférentielle** qui modélisent les préférences de l'utilisateur. Pour mesurer la santé d'une entreprise, la direction financière accorde probablement plus d'importance aux chiffres d'affaires et de ventes alors que la direction des personnels

accorde probablement plus d'importance à des indicateurs humains comme le nombre d'arrêts maladie par exemple.

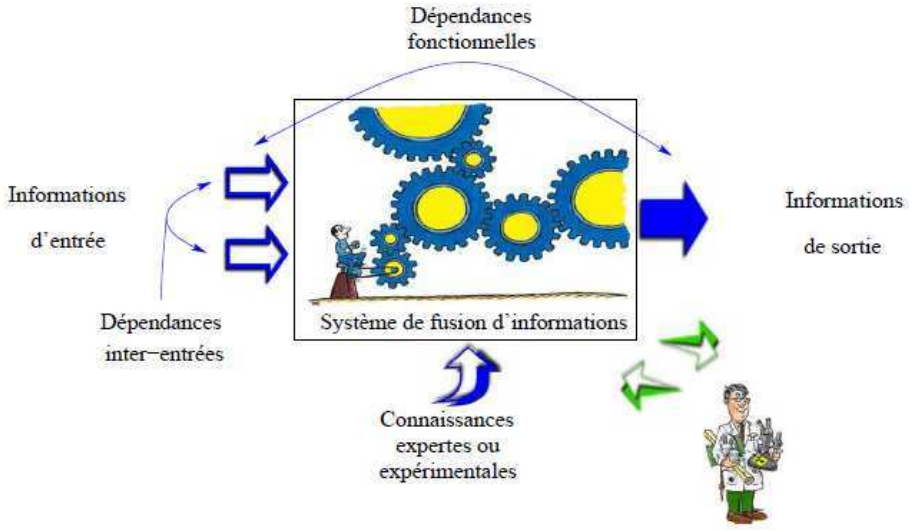



FIGURE 5.3 – Les dépendances dans un système de fusion d'information) ISSU DE [JULLIEN, 2008]

Au niveau de l'organisation de cette étape de combinaison on retrouve différentes architectures. Ces architectures de combinaison d'informations peuvent être classées suivant deux catégories lorsque l'on s'intéresse à l'aspect temporel de la fusion [Valet, 2001] :

**L'organisation parallèle** est l'architecture la plus utilisée, les informations sont disponibles en même temps et sont agrégées en même temps.

**L'organisation série** consiste à agréger les informations successivement. Les informations d'entrée sont ordonnées : les deux premières sont agrégées, puis ce résultat intermédiaire est agrégé avec la troisième entrée et ainsi de suite.

L'implémentation des opérateurs d'agrégation repose sur un cadre mathématique bien défini. En effet, son rôle est double puisqu'il permet la représentation des informations dans un espace commun et permet l'agrégation par l'opérateur de fusion de ces informations. De plus, la modélisation de l'incertitude et de l'imprécision des informations traitées est directement liée au formalisme mathématique choisi. Parmi les cadres les plus classiques, on trouve la théorie des probabilités, la théorie des possibilités et la théorie de l'évidence [Rombaut, 2001]. Nous n'allons pas détailler ces théories car largement traitées dans la littérature mais nous allons présenter succinctement les fondements et quelques opérateurs de combinaisons.

#### 5.1.5.1 Théorie des probabilités

Les probabilités offrent le plus ancien formalisme et mettent à la disposition de l'utilisateur un certain nombre d'outils mathématiques qui lui permettent de régler la majorité des

problèmes rencontrés surtout lorsque l'on a une approche statistique du problème à traiter (ce qui n'est pas toujours le cas). Grâce à ce formalisme les informations d'entrée sont modélisées dans un espace commun où l'incertitude et l'imprécision peuvent être représentées à l'aide de probabilités ou de distributions de probabilités (où le modèle Gaussien est souvent utilisé). L'inférence Baysienne qui se décline dans le cas continu et dans le cas discret est la méthode privilégiée de combinaison dans la théorie des probabilités. Il s'agit pour une information conditionnée par toute l'information disponible d'évaluer sa probabilité d'être vraie. Supposons que l'on dispose de $n$ observations $s_i$, avec $i \in [1;n]$, pour estimer dans quelle mesure des hypothèses $H_j$ sont vraies. Pour chaque source d'information, on dispose des probabilités conditionnelles $P(s_i|H_j)$ modélisant l'incertitude sur les mesures, c'est-à-dire la probabilité d'avoir la mesure $s_i$ sachant que l'hypothèse $H_j$ est vraie. On suppose aussi que l'on dispose de la probabilité *a priori* $P(H_j)$ sur les hypothèses. Si les sources d'information sont indépendantes alors la relation de Bayes est donnée par l'equation 5.4.

$$P(H_j|s_i) = \frac{P(H_j)\prod_i P(s_i|H_j)}{\sum_k P(H_k)\prod_i P(s_i|H_k)} \tag{5.4}$$

Les probabilités $P(s_i|H_j)$ et $P(H_j)$ sont en pratique rarement connues. Elles sont souvent estimées à partir des données ou connues par expérience.

#### 5.1.5.2 Théorie de l'évidence

La théorie de l'évidence, appelée aussi théorie de la croyance ou théorie de Dempster Shafer, initiée par ces deux auteurs [Dempster, 1968, Shafer, 1976], est relativement récente. C'est une généralisation de l'inférence bayésienne au traitement de l'incertain. Elle permet de manipuler des événements non nécessairement exclusifs. Cette capacité lui confère l'avantage de pouvoir représenter explicitement, par l'utilisation de degrés de croyance, l'incertitude sur un événement.

**5.1.5.2.1 Cadre de la théorie** Dans la théorie de l'évidence, le raisonnement porte sur le cadre de discernement $\Omega$ défini comme un ensemble de $N$ hypothèses $H_i$ exclusives et exhaustives. L'ensemble noté $2^\Omega$ sert de référentiel de définition pour évaluer la véracité d'une proposition $A$. Cette proposition peut par exemple être l'ensemble $A = \{H_1 \cup H_2 \cup H_3\}$ notée le plus souvent $A = \{H_1, H_2, H_3\}$. Le référentiel de définition $2^\Omega = \{\varnothing, H_1, ..., H_N, H_1 \cup H_2, \ldots, H_1 \cup H_2 \cup H_3, \ldots, \Omega\}$ est composé de l'ensemble des $2^N$ **sous-ensembles** $A$ de $\Omega$.

Le formalisme mathématique de cette théorie repose tout d'abord sur la définition de masses accordées aux événements. Pour exprimer un degré de confiance pour chaque proposition $A$ de $2^\Omega$, il est possible de lui associer une masse d'évidence élémentaire $m(A)$ qui indique toute la confiance que l'on peut avoir dans cette proposition sans pour autant privilégier aucune des hypothèses qui la composent. Cette masse $m(A)$ correspond au degré de croyance placée **exactement** sur la proposition A. Si A n'est pas une hypothèse singleton (hypothèse simple $H_i$ et non une disjonction d'hypothèses) alors cette masse ne peut, compte tenu de l'état actuel de la connaissance, être affectée à un sous-ensemble plus spécifique de $A$. La masse $m(A)$ affectée à une disjonction d'hypothèses $A$ est vue comme toute la masse susceptible d'être transférée ultérieurement à un sous-ensemble plus spécifique à cette disjonction (sous réserve d'apports d'informations supplémentaires permis par la loi de combinaison de

Dempster). On l'appelle alors masse potentielle pour chaque hypothèse participant à cette disjonction. La fonction $m$ est définie de $2^\Omega$ sur $[0;1]$ par :

$$\begin{aligned} m : 2^\Omega &\rightarrow [0;1] \\ A &\rightarrow m(A) \end{aligned}$$

et vérifie les propriétés :

$$\begin{aligned} m(\varnothing) &= 0 \\ \sum_{A \subseteq \Omega} m(A) &= 1 \end{aligned}$$

Tout $A \subseteq \Omega$ avec $m(A) > 0$ est appelé élément focal de $2^\Omega$. L'ensemble des éléments focaux constitue le noyau $N_\Omega$. Notons que lorsque $N_\Omega = \Omega$ la notion de masse élémentaire est assimilable à celle de probabilité. De plus, un apport du modèle de l'évidence par rapport à l'approche probabiliste est de ne pas être obligé de répartir la masse totale de probabilité sur des singletons permettant ainsi d'avoir une attitude moins arbitraire. En effet, affecter une masse non nulle à une proposition $A$ qui n'est pas un singleton, indique que l'ensemble des hypothèses de $A$ nous paraît crédible mais sans pour autant prendre parti particulièrement pour l'une d'entre elles [Chauveau, 2009].

**5.1.5.2.2 Règle de combinaison de Dempster** La théorie de l'évidence offre des outils appropriés pour la fusion de sources d'informations incertaines et/ou imprécises. A partir des jeux de masses notés $m_{S_k}$ obtenus sur chacune des $M$ sources d'information $S_k$, il est possible de construire un jeu de masses unique $m$ par simple sommation orthogonale des jeux de masses $m_{S_k}$ en utilisant une règle de combinaison comme celle de Dempster. Ce jeu de masses $m$ synthétise toute la connaissance contenue dans les jeux de masses issus de chacune des différentes sources et peut alors être utilisé par un module de décision. Historiquement, l'opérateur de Dempster (appelé également somme orthogonale) est le premier opérateur de combinaison défini dans le cadre de la théorie de l'évidence. Son utilisation impose de respecter la condition d'indépendance des sources d'information à combiner. La masse résultant de la combinaison de $M$ sources d'information $S_k$ est notée $m_\oplus$ et est définie comme ceci [Lefevre *et al.*, 2001] :

$$m_\oplus = m_{S_1} \oplus \ldots \oplus m_{S_k} \oplus \ldots \oplus m_{S_M}$$

où $\oplus$ représente l'opérateur de combinaison de Dempster et s'écrit :

$$m_\oplus(A) = \frac{m_\cap(A)}{1-K} = \frac{\sum_{A_1 \cap \ldots \cap A_M \neq \varnothing} \{\prod_{k=1}^{M} m_{S_k}(A_k)\}}{1 - \sum_{A_1 \cap \ldots \cap A_M = \varnothing} \{\prod_{k=1}^{M} m_{S_k}(A_k)\}}$$

où le terme $m_\cap(A)$ correspond à la règle de combinaison conjonctive et où $K$ qui représente la masse affectée à l'ensemble vide traduit le conflit existant entre les sources. Lorsque

ce coefficient est égal à 1, les sources sont en conflit total et ne peuvent être fusionnées. A l'inverse, lorsque ce coefficient est égal à 0, les sources sont en accord parfait.

De nombreuses applications utilisent la théorie des fonctions de croyance comme cadre mathématique pour fusionner des informations dans des domaines variés comme en imagerie sonar [Maussang, 2005] ou dans l'analyse des séquences sportives d'athlétisme [Ramasso, 2007]. L'état de l'art de [Ramasso, 2007] propose une vision plus détaillée des combinaisons possibles dans cette théorie.

#### 5.1.5.3 Théorie des possibilités et des ensembles flous

Cette théorie récente associée aux ensembles flous, introduite dans les années 70 par Zadeh [Zadeh, 1975] puis développée par Dubois et Prade [Dubois et Prade, 1988], constitue un cadre permettant de traiter les concepts d'imprécision et d'incertitude de nature non probabiliste. En effet, elle fournit le moyen de dire dans quelle mesure la réalisation d'un événement est possible et dans quelle mesure on en est certain, sans toutefois avoir à disposition l'évaluation de la probabilité de cette réalisation.

Étant donné un ensemble de référence fini $\Omega$, on attribue à chaque événement $A$ défini sur $\Omega$, c'est-à-dire à chaque sous ensemble $A$ de $\Omega$ un coefficient compris entre 0 et 1 évaluant à quel point cet événement est possible. Dans cette théorie, ce coefficient correspond à la *mesure de possibilité* $\Pi$ qui est une fonction définie sur l'ensemble $2^\Omega$ des parties de $\Omega$ :

$$\begin{aligned} \Pi : 2^\Omega &\rightarrow [0;1] \\ A &\rightarrow \Pi(A) \end{aligned}$$

et vérifie les propriétés suivantes :

$$\begin{aligned} \Pi(\varnothing) &= 0 \\ \Pi(\Omega) &= 1 \end{aligned}$$

et

$$\Pi(\cup_{i=1,2,\ldots} A_i) = sup_{i=1,2,\ldots}(\Pi(A_i))$$

où *sup* indique le supremum des valeurs concernées, soit la plus grande d'entre elles dans le cas fini. Soient $A1$ et $A2$ deux événements de $2^\Omega$ alors on définit :

$$\Pi(A1 \cup A2) = max(\Pi(A1), \Pi(A2))$$

Une mesure de possibilité est totalement définie si l'on attribue un coefficient de possibilité à toute partie de l'ensemble de référence $\Omega$. Elle est définie plus simplement si l'on indique

les coefficients attribués seulement aux parties élémentaires de $\Omega$, une partie quelconque étant l'union de parties élémentaires. Une fonction de *distribution de possibilité* $\pi$ permet d'attribuer un degré de possibilité à tout élément de $\Omega$ et non plus à toute partie de $\Omega$ :

$$\begin{aligned} \pi : \Omega &\rightarrow [0;1] \\ H_i &\rightarrow \pi(H_i) \end{aligned}$$

et

$$\forall A \in 2^{\Omega} \quad \Pi(A) = sup_{H_i \in A}(\pi(H_i))$$

La théorie des possibilités permet le traitement d'incertitudes de nature non probabiliste sur des événements décrits sans imprécision ni caractéristique vague. Mais historiquement, Zadeh a introduit la théorie des possibilités à propos de la caractérisation de variables par des descriptions linguistiques imprécises, représentées par des sous-ensembles flous. Un ensemble flou $A$ de $X$ est défini par une fonction d'appartenance $\mu$ qui associe à chaque élément $x$ de $X$, le degré $\mu_A(x)$, compris entre 0 et 1, avec lequel $x$ appartient à $A$. La fonction d'appartenance de ces ensembles flous conduit à la définition d'une distribution de possibilités, qui permet de traiter les incertitudes engendrées au cours d'un raisonnement fondé sur les caractérisations floues des variables. En effet, une caractérisation floue telle que « grand » est définie *a priori* et sa fonction d'appartenance $\mu_A$ indique avec quel degré chaque élément de $X$ lui appartient. Une proposition floue telle que « la taille est grande » est une description floue *a posteriori* de la variable linguistique « taille », après observation d'une situation particulière, qui décrit de façon vague la taille d'un individu donné et indique dans quelle mesure il est possible que sa taille exacte soit tel ou tel élément de $X$. Ceci veut dire qu'une proposition floue induit une *distribution de possibilité* $\pi_{V,A}$ sur $X$, définie à partir de la fonction d'appartenance associée à $A$ par [Bouchon-Meunier, 1993] :

$$\forall x \in X \quad \pi_{V,A}(x) = \mu_A(x)$$

Cette définition exprime que, si $\varepsilon$ est le degré d'appartenance d'un élément quelconque $x$ de $X$ à la caractérisation floue $A$, la possibilité pour que la variable $V$ prenne la valeur $x$, sachant que $V$ est caractérisé par $A$, est aussi égale à $\varepsilon$.

**5.1.5.3.1 Les principaux opérateurs d'agrégation** Le fait d'utiliser des sous-ensembles flous pour décrire des classes imparfaitement localisées dans $X$, conduit à caractériser, par exemple, les points de $X$ communs à différentes classes ou bien étrangers à ces classes. Les notions d'inclusion, d'intersection, d'union, de complément de sous-ensembles flous sont donc utiles. Ces opérations sur les sous-ensembles flous sont effectuées à l'aide d'opérateur qui peuvent être regroupés en quatre grandes classes [Grabisch et Perny, 2001] :

**Les opérateurs conjonctifs** effectuent une agrégation des quantités comme le ferait un "ET" logique (conjonction). Ainsi le résultat de l'agrégation est élevé (proche de 1) si et seulement si toutes les quantités à agréger sont élevées [Grabisch et Perny, 2001]. La famille des normes triangulaires ou t-normes, souvent notées $\top$ est un sous ensemble de

ces opérateurs. Ses caractéristiques sont bien connues (commutativité, associativité, 1 est l'élément neutre, 0 est l'élément absorbant, monotonie) [Fodor et Roubens, 1994]. Parmi les principales t-normes citons :

- **t-norme de Zadeh** :

$$\top(x, y) = min(x, y)$$

- **t-norme probabiliste** :

$$\top(x, y) = xy$$

- **t-norme de Lukasiewicz** :

$$\top(x, y) = max(x + y - 1, 0)$$

- **t-norme de Weber** :

$$\begin{cases} \top(1, y) = y \\ \top(x, 1) = x \\ \top(x, y) = 0 \quad \text{sinon} \end{cases}$$

**Les opérateurs disjonctifs** effectuent une agrégation de type "OU" logique (disjonction). Le résultat de l'agrégation est élevé dès que l'une des quantités à agréger est élevée. La famille des co-normes triangulaires ou t-conormes, souvent notées $\perp$ est un sous ensemble de ces opérateurs dont les caractéristiques sont bien connues (commutativité, associativité, 0 est l'élément neutre, 1 est l'élément absorbant, monotonie). Parmi les principales t-conormes citons :

- **t-conorme de Zadeh** :

$$\perp(x, y) = max(x, y)$$

- **t-conorme probabiliste** :

$$\perp(x, y) = x + y - xy$$

- **t-conorme de Lukasiewicz** :

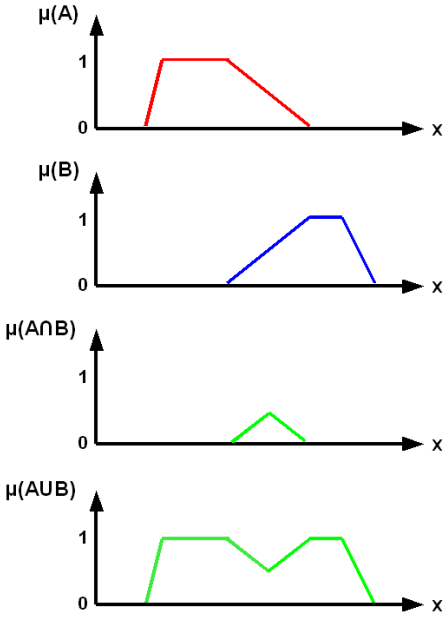
$$\perp(x, y) = min(x + y, 1)$$

- **t-conorme de Weber** :

$$\begin{cases} \perp(0, y) = y \\ \perp(x, 0) = x \\ \perp(x, y) = 1 \quad \text{sinon} \end{cases}$$

**Les opérateurs de compromis** se situent par définition entre les opérateurs disjonctifs et conjonctifs. Ce sont par exemple la somme pondérée, les opérateurs de moyenne, le minimum et maximum pondérés, les intégrales floues comme celle de Sugeno [Sugeno, 1974]. Voir [Grabisch et Perny, 2001] pour le détail de ces opérateurs.

**Les opérateurs hybrides** sont les opérateurs qui ne peuvent être classés dans les catégories précédentes, comme par exemple les opérateurs de Zimmermann et Zysno qui sont un mélange de t-normes et de t-conormes [Zimmermann et Zysno, 1980].




La conjonction (**ET** logique) de Zadeh de l'ensemble flou A et de l'ensemble flou B définis sur l'univers de discours U est un ensemble flou de fonction d'appartenance :

$$\mu_{A\cap B}(x)) = min(\mu_A(x), \mu_B(x)) \ \forall x \in U$$

La disjonction (**OU** logique) de Zadeh de l'ensemble flou A et de l'ensemble flou B définis sur l'univers de discours U est un ensemble flou de fonction d'appartenance :

$$\mu_{A\cup B}(x)) = max(\mu_A(x), \mu_B(x)) \ \forall x \in U$$

FIGURE 5.4 – Exemple d'opérateurs flous conjonctifs et disjonctifs

**5.1.5.3.2 Les systèmes flous** Les sous-ensembles flous et la théorie des possibilités sont des éléments importants de la représentation des connaissances imparfaitement définies. Pour raisonner sur de telles connaissances (imprécises, vagues et/ou incertaines) on utilise la logique floue dont la mise en œuvre peut être réalisée dans les systèmes flous. Un système flou (ou contrôleur flou dans le domaine de l'automatique) peut être vu comme un système expert simple et fonctionnant à partir d'une représentation des connaissances basée sur les ensembles flous [Bouchon-Meunier, 1993]. D'une manière générale, un système flou est constitué de trois étapes (voir figure 5.5) :

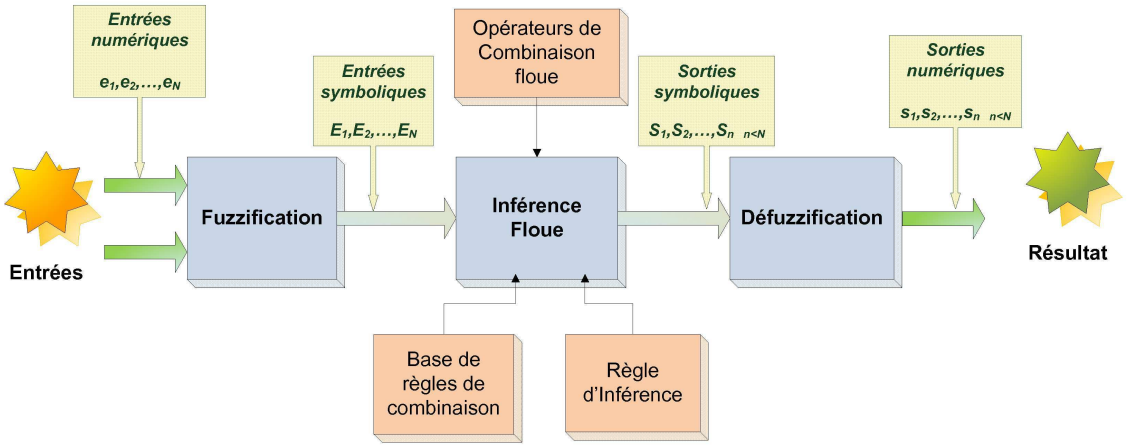



FIGURE 5.5 – Configuration générale d'un système flou

1. La **fuzzification** réalise l'interface avec le monde extérieur. Elle consiste à exprimer linguistiquement les valeurs numériques d'entrée de nature différentes et vise à les exprimer dans un espace homogène commun ce qui permet ainsi d'agréger des données

hétérogènes. Cette étape de fuzzification correspond à l'étape de représentation de l'information de la figure 5.1.

2. L'**inférence** ou raisonnement flou est l'étape d'agrégation des données et correspond à l'étape de combinaison dans la décomposition des systèmes de fusion (voir figure 5.1). Elle vise à transformer la partie floue $E_n$ issue de la fuzzification en une nouvelle partie floue $S_n$. La théorie des sous-ensembles flous propose divers mécanismes pour réaliser cette combinaison dont le plus usité est la règle floue. Ces règles sont consignées dans une base de règles de type SI-ALORS et traduisent symboliquement la connaissance des experts. La combinaison des entrées floues est réalisée grâce aux opérateurs de combinaison floue puis l'utilisation d'une règle d'inférence floue permet d'obtenir la sortie symbolique.

3. La **défuzzification** consiste à transformer si nécessaire le résultat flou inféré pour le mettre sous une forme nette (sous forme numérique par exemple dans le cas d'un contrôleur flou). Cette étape de défuzzification correspond à l'étape d'interprétation de l'information de la figure 5.1.

La réalisation d'un système flou est recommandée lorsque les règles de fusion sont mal définies ou difficiles à décrire précisément, par exemple en raison d'une trop grande complexité. Elle est également très utile lorsque les variables intervenant dans le processus sont caractérisées de façon imprécise ou lorsque les connaissances sont exprimées en langage naturel et non numériquement.

### 5.1.6 L'interprétation de l'information

Cette dernière étape consiste à transformer l'information obtenue en sortie de l'étape de combinaison sous une forme exploitable pour son utilisation par un organe de commande ou un utilisateur humain. Cela consiste à changer la représentation de cette information directement par une transformation mathématique, ou à l'aide de la connaissance d'un expert. Par exemple, dans le cas d'une régulation lorsque l'information fusionnée est sous la forme d'une distribution de probabilités ou de possibilités, l'étape d'interprétation permet d'obtenir la valeur numérique de la commande (elle est alors assimilable à l'opération de défuzzification de la figure 5.5). Cette étape sert généralement à la prise de décision mais peut également consister en l'évaluation de l'information fusionnée pour permettre d'optimiser le système de fusion par un bouclage de cette information interprétée sur les paramétres de réglage des étapes précédentes.

## 5.2 Présentation des objectifs et de la méthodologie de fusion

Nous avons vu dans les chapitres précédents que notre objectif est la caractérisation des films d'animation à partir de sources hétérogènes que sont les images et le texte. Les informations de couleur, de rythme et d'activité sont extraites des séquences d'images séparément des informations textuelles comme la description de l'action et de l'atmosphère caractérisant le synopsis. Cette caractérisation des films d'animation, comme nous l'avons vu au chapitre sur l'analyse des textes (§4.5), peut être réalisée à deux niveaux différents :

- La fusion de ces informations peut servir pour caractériser le film de façon **globale**. Cette caractérisation globale est envisagée dans ce chapitre à travers le **genre d'animation**. En effet, ce champ est très souvent mal ou pas renseigné dans les fiches d'inscription au festival. Les informations apportées par l'image comme la couleur et ses contrastes peuvent traduire une atmosphère dans la séquence vidéo qui peut être complétée et donc fusionnée par une atmosphère traduite par le texte du synopsis. Une deuxième façon de caractériser la séquence d'animation au niveau global passe par la caractérisation de son **activité**. Là encore les indicateurs image et texte sont des caractéristiques pertinentes pour caractériser l'activité de la séquence vidéo.

- La fusion de ces informations peut servir également pour caractériser le film de façon **locale**. La deuxième façon de caractériser la séquence d'animation est envisagée ici au **niveau local** c'est-à-dire située dans le temps. En effet, nous avons vu avec le *scénario actanciel* que les synopsis pouvaient être une description d'une ou de plusieurs partie(s) du film où une action et des actants étaient mis en scène. L'idée est donc d'aligner nos deux sources d'information que sont les images et le texte pour pouvoir décrire localement les éléments constituant la sous-séquence.

Pour atteindre ces objectifs de caractérisation à deux niveaux nous proposons deux approches pour combiner les informations issues de l'image et du texte. Nous présentons ci après les éléments et le cadre méthodologique de ces approches.

- Une approche par règles expertes. Cette approche de fusion d'information est basée sur l'intégration de connaissances fournies par un expert. Cette connaissance est codée par des règles de fusion qui traduisent les dépendances entre les informations de sortie et d'entrée du système de fusion. La théorie mathématique utilisée pour représenter et combiner ces informations de nature différente (numérique et symbolique) est la théorie des sous-ensembles flous. En effet, cette théorie propose un cadre adapté pour coder la connaissance des experts sous forme de règles descriptives claires et compréhensibles. Ainsi, les professionnels de l'animation peuvent comprendre et appréhender plus facilement le processus de combinaison des informations qui conduit à caractériser automatiquement ces films. Cet aspect est important car il n'est pas facile pour un artiste d'imaginer et d'accepter qu'un programme puisse analyser automatiquement son œuvre. Dans ce sens, les informations numériques d'entrée peuvent être évaluées par des mots attachés à des concepts au moyen des descriptions linguistiques. Ces descriptions sont obtenues à l'aide de fonctions d'appartenance permettant le passage de l'univers numérique à l'univers symbolique. Le choix de cette approche symbolique est renforcé par le fait que l'univers de sortie est non numérique (caractérisation symbolique du film).

- Une approche par apprentissage. Cette approche de fusion d'information est basée sur l'utilisation d'algorithmes de classification supervisée qui permettent d'exhiber automatiquement les liens fonctionnels entre les informations d'entrées et de sortie. Cette approche est utilisée lorsque l'on ne dispose pas de connaissances *a priori* sur ces relations qui sont obtenues automatiquement par apprentissage à partir d'une base d'exemples.

## 5.3 Caractérisation globale des films appliquée au genre des films d'animation

Notre objectif dans cette section est de caractériser d'un point de vue global le film à travers une information qui est le genre des films d'animation. Ce choix a plusieurs motivations :

- Lorsque l'on regarde la répartition des genres déclarés des films d'animation dans la base des fiches d'inscription au festival (voir figure B.6 de l'annexe B), on remarque que dans 25% des cas les genres ne sont pas renseignés. Dans les autres cas cette information importante est souvent mal renseignée. Par conséquent, la caractérisation automatique des films permettra de compléter ou de corriger ce champ.

- nous disposons d'une ontologie des genres créée avec l'aide des experts du cinéma d'animation permettant de s'appuyer sur des concepts hiérarchisées et des relations entre les différents genres [Beauchêne et Deloule, 2009].

- les informations images et textuelles traduisent une atmosphère qui peut être liée au genre du film. En effet les couleurs et les mots utilisés dans ces œuvres sont issus d'un choix artistique et traduisent l'ambiance ou l'atmosphère voulue par l'auteur. Ce lien entre la couleur et l'atmosphère dégagée par l'image est fort et bien connu du monde artistique (voir §2.1.2.2).

Dans cette étude sur le rapprochement entre le texte et l'image nous avons fait le choix de nous focaliser dans un premier temps sur le genre **"drame"**. En effet, le lien entre la couleur et l'atmosphère dégagée par l'image est renforcé lorsque le film aborde des sujets graves et dramatiques. De plus, le manque de diversité des couleurs ainsi que l'utilisation abondante de couleurs froides ou sombres (descripteurs images dont nous disposons) marquent cette volonté de plonger le spectateur dans un contexte noir. Cependant ce lien entre le drame et les informations de couleurs peut être **incertain**. En effet, un film humoristique peut se passer la nuit sans que le sujet abordé soit dramatique. L'auteur peut dans une volonté artistique vouloir se détacher de la « norme » et jouer avec les genres en traitant par exemple un sujet dramatique à l'aide d'images colorées, chaudes et aux couleurs variées. Ainsi, l'utilisation d'une information complémentaire d'un plus haut niveau sémantique comme le texte (mesure de l'intensité thématique du drame) peut compléter ces informations apportées par l'image. Par conséquent, la fusion des informations **complémentaires** issues des **images** et du **texte** a pour objectif d'apporter une information de meilleure qualité en diminuant son **incertitude** afin d'arriver à prendre une **décision** quant à l'appartenance de la séquence d'animation au genre **dramatique**.

La figure 5.6 présente le système de fusion développé dans ce but. La caractérisation d'un film d'animation à partir des informations issues du texte et de l'image (à droite sur la figure) est effectuée à travers trois étages. Un premier étage (qui peut être relié à l'étape de représentation dans la décomposition des systèmes de fusion (voir figure 5.1)) permet de combiner les informations numériques issues de chaque source d'information afin d'obtenir de nouveaux concepts (froideur, monotonie et uniformité) qui traduisent la volonté artistique de plonger le spectateur dans une atmosphère dramatique. Ces nouvelles informations symboliques sont ensuite fusionnées dans un deuxième étage (qui peut être relié à l'étape de

combinaison dans la décomposition des systèmes de fusion (voir figure 5.1)) dédié à chacune des sources afin d'obtenir une information liée au drame. Ces deux informations issues du texte et de l'image (notées (7) et (8)) indiquent respectivement la possibilité que le synopsis et les images traduisent une atmosphère dramatique. Finalement le dernier étage (qui peut également être relié à l'étape de combinaison dans la décomposition des systèmes de fusion (voir figure 5.1)) permet de fusionner ces deux informations complémentaires pour donner une information symbolique globale et certaine [1] sur l'atmosphère dramatique que dégage le film d'animation.

1. c'est-à-dire dont l'incertitude a été diminuée

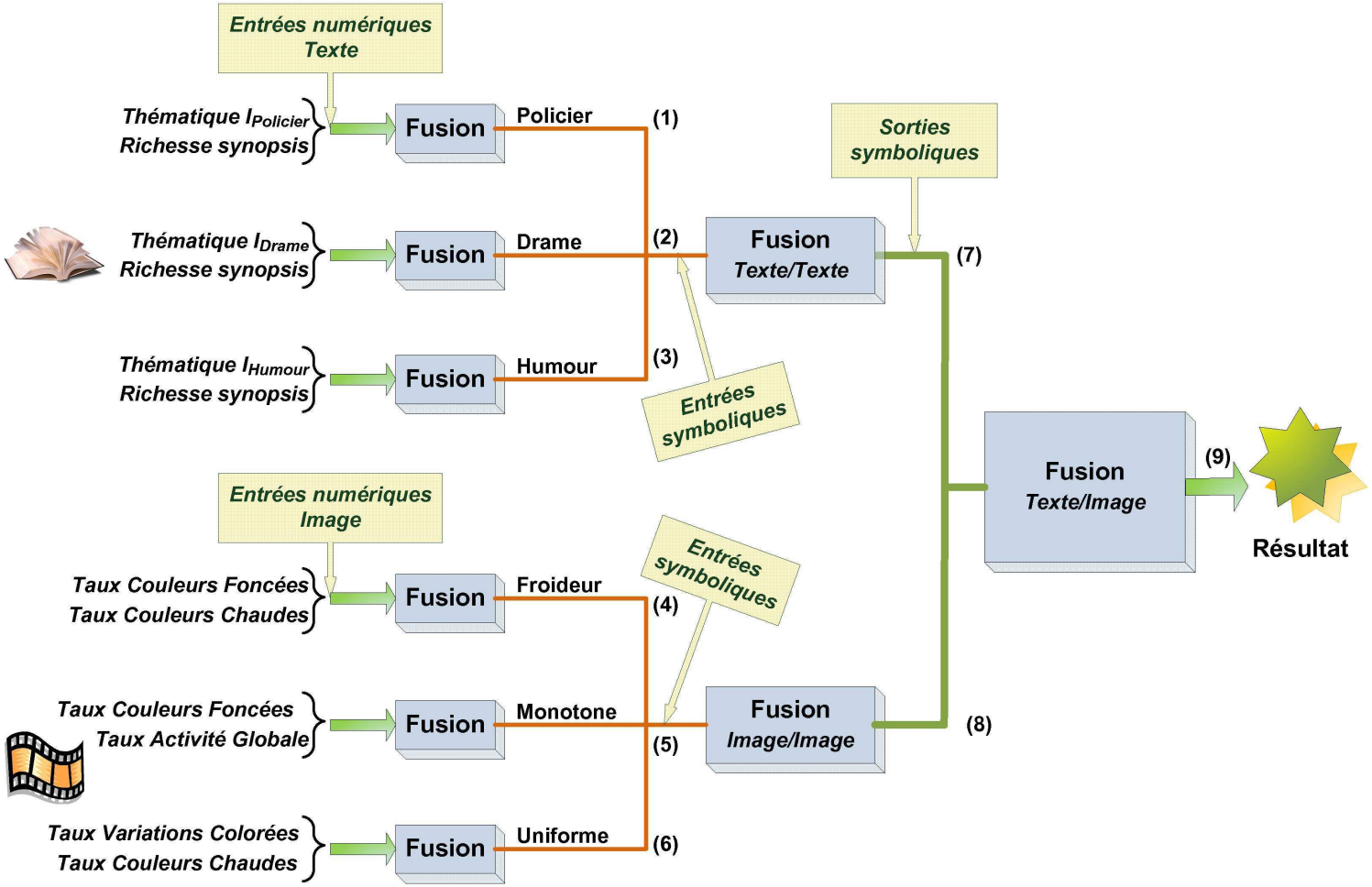



FIGURE 5.6 – Structure du système de fusion d'information pour la caractérisation du genre dramatique

### 5.3.1 Fusion des indicateurs texte

Nous avons vu dans le chapitre précédent une méthode d'analyse thématique des synopsis pour identifier une atmosphère liée au drame. Nous avons conclu que la mesure de l'intensité lexicale seule n'était pas suffisante pour retrouver tous les films dont le genre déclaré est le drame (rappel 34%). De plus, avec cette mesure, des genres comme "policier", "humour" ou "satire" sont souvent confondus avec le genre dramatique (précision de 24%). Par conséquent il est nécessaire d'améliorer cette information textuelle avant de la fusionner à l'information image. Pour cela nous utilisons trois informations supplémentaires issues de l'analyse des textes :

**L'intensité thématique liée au genre "Policier"** est utilisée pour diminuer l'incertitude quant à la mesure du drame et ainsi permettre d'augmenter la mesure de précision de la recherche du drame. Voir annexe D.3.2 pour le détail du calcul de cet indicateur.

**L'intensité thématique liée au genre "Humour"** est utilisée pour diminuer l'incertitude quant à la mesure du drame et ainsi permettre d'augmenter la mesure de précision de la recherche du drame. Voir annexe D.3.3 pour le détail du calcul de cet indicateur.

**La richesse du synopsis** permet de nuancer la mesure de l'intensité thématique. Les synopsis des films d'animation ne font pas tous la même longueur et il est important de prendre en compte cette information. En effet, plus un texte est long plus le nombre de mots « parasites » par rapport à la thématique recherchée est important (il y a par exemple plus de qualificatifs dans la phrase), l'intensité thématique est donc moins forte. A l'inverse plus le texte est court plus l'information est synthétique et plus l'intensité thématique est forte. Par exemple dans la phrase « Cette histoire est dramatique » où il y a peu de qualificatifs, le terme dramatique (qui appartient au dictionnaire du drame) apparaît une seule fois et a un fort impact du fait de la brièveté du texte (4 mots). Par conséquent l'intensité thématique du drame pour cette phrase sera choisie égale à 25% (1/4). Par contre dans la phrase « Dans ce petit conte pour enfants, l'histoire est dramatique » l'intensité thématique du drame ne vaudra plus que 10%[2]. Ainsi nous définissons la **richesse** d'un texte comme le nombre de termes différents le composant. Cette mesure permet de nuancer la mesure de l'intensité thématique.

Dans un premier temps nous fusionnons les deux sources d'informations textuelles que sont l'**intensité thématique** et la **richesse** du synopsis (premier système de fusion en haut à gauche sur la figure 5.6). Nous optons pour un système d'agrégation floue avec l'utilisation de règles de combinaison issues de l'expertise du domaine. Le mécanisme de combinaison de ces informations est expliqué après.

#### 5.3.1.1 La fuzzification

Les informations textuelles retenues pour caractériser l'atmosphère sont des mesures statistiques purement numériques (occurrences et nombre de mots). Elles sont donc transformées

2. Afin de simplifier les calculs de cet exemple on ne lemmatise pas le texte et on ne supprime pas les mots outils

en valeurs symboliques par l'utilisation d'ensembles flous.

Ainsi le concept de possibilité du genre dramatique associé à l'intensité thématique du même nom $I_{Drame}$ est décrit en utilisant cinq variables linguistiques illustrées par les symboles suivants : *possibilité* ***T**rès **F**aible d'être du Drame*, *possibilité* ***F**aible d'être du Drame*, *possibilité* ***M**oyenne d'être du Drame*, *possibilité* ***H**aute d'être du Drame* et *possibilité* ***T**rès **H**aute d'être du Drame*. Ce partitionnement de l'univers de discours $I_{Drame}$ est noté $L_{Drame}(I_{Drame})\{TF, F, M, H, TH\}$ et il est choisi afin de représenter « au mieux » (suivant l'expertise) les différentes classes sémantique possibles de cette variable linguistique. Ce choix du nombre et la signification floue des symboles est obtenu par expertise du domaine. La signification floue de chaque symbole (ou terme linguistique) correspond à déterminer le sous-ensemble flou des nombres qu'il représente. Elle est illustrée sur la figure 5.7 par sa fonction d'appartenance ($\mu_{TF}$, $\mu_F$, $\mu_M$, $\mu_H$ et $\mu_{TH}$) de type trapézoïdale.

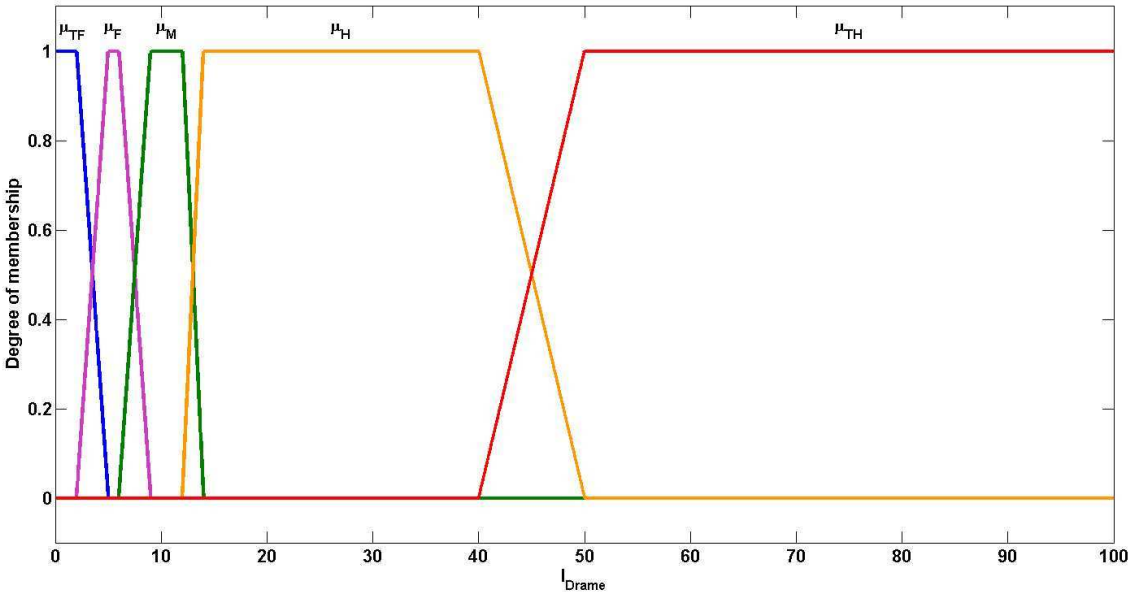



FIGURE 5.7 – La signification floue des cinq termes du partitionnement $L_{Drame}$ de l'univers de discours de l'intensité dramatique $I_{Drame}$ est déterminée par les fonctions d'appartenance floues : $\mu_{TF}(I_{Drame}) = 1, \forall I_{Drame} \in [0, 2]$, $\mu_F(I_{Drame}) = 1, \forall I_{Drame} \in [5, 6]$, $\mu_M(I_{Drame}) = 1, \forall I_{Drame} \in [9, 12]$, $\mu_H(I_{Drame}) = 1, \forall I_{Drame} \in [14, 40]$ et $\mu_{TH}(I_{Drame}) = 1, \forall I_{Drame} \in [50, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

Le concept de longueur du texte associé à la mesure de richesse est décrit en utilisant également cinq variables linguistiques illustrées par les symboles suivants : *synopsis Très Court*, *synopsis Court*, *synopsis Moyen*, *synopsis Long* et *synopsis Très Long*. La signification floue de chaque symbole de $L_{richesse}$ est traduite par sa fonction d'appartenance (de type trapézoïdale) : $\mu_{TC}$, $\mu_C$, $\mu_M$, $\mu_L$ et $\mu_{TL}$, et est illustrée par la figure 5.8.

De plus, et de façon générale, le partitionnement $L(X) = \{A_1, A_2, \ldots, A_n\}$ des univers de discours $X$ que nous traitons dans ce travail est réalisé de façon à avoir :

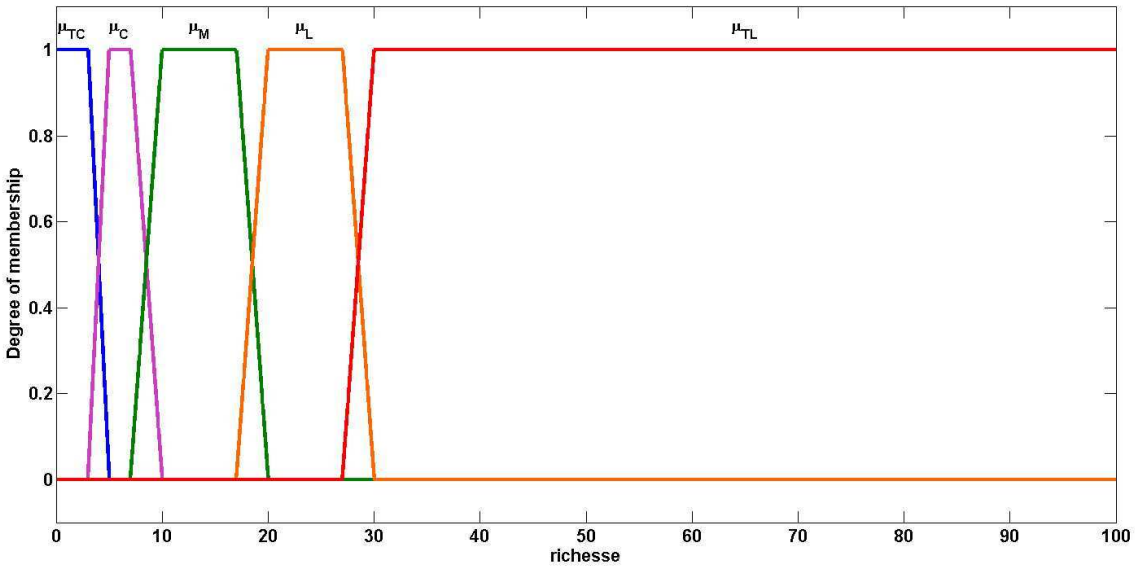



FIGURE 5.8 – La partition floue de l'univers de discours de la richesse d'un synopsis est déterminée par l'ensemble des fonctions d'appartenance floues : $\mu_{TC}(richesse) = 1, \forall richesse \in [0,3]$, $\mu_C(richesse) = 1, \forall richesse \in [5,7]$, $\mu_M(richesse) = 1, \forall richesse \in [10,17]$, $\mu_L(richesse) = 1, \forall richesse \in [20,27]$ et $\mu_{TL}(richesse) = 1, \forall richesse \in [30,100]$, (l'axe des ordonnées correspond au degré d'appartenance).

$$\forall x \in X \sum_{A_i \in L(X)} \mu_{A_i}(x) = 1 \tag{5.5}$$

Finalement, la description floue d'une information (ici une valeur numérique d'entrée) consiste à déterminer le sous-ensemble des termes linguistiques qui la qualifie. Elle est obtenue en calculant le degré d'appartenance de cette grandeur numérique à chacun des symboles servant à d'écrire l'attribut. Par exemple si la valeur numérique de l'intensité dramatique vaut 8 alors la description floue $F_{Drame}$ de cette entrée est la suivante :

$$F_{Drame}(8) = 0/\text{Très Faible} + 0.33/\text{Faible} + 0.67/\text{Moyen} + 0/\text{Haute} + 0/\text{Très Haute}$$

où la description floue $F_{Drame}$ de cette entrée est composée d'un degré d'appartenance de 0.33 au symbole **Faible** et d'un degré d'appartenance de 0.67 au symbole **Moyen**. De plus, la contrainte donnée dans l'équation 5.5 correspond à un mode de raisonnement probabiliste beaucoup plus compréhensible et proche du raisonnement des experts.

#### 5.3.1.2 La base de règles

La fusion de ces deux informations symboliques se fait en utilisant un ensemble de règles Si-Alors ("IF-THEN") obtenues par expertise et qui permettent d'obtenir une information caractérisant la possibilité pour un synopsis d'être associé à une atmosphère ou un genre dramatique. Une règle se compose de prémisses et d'une conclusion et représente les relations et combinaisons entre les entrées et la sortie.

Par exemple, dans la règle :

**Si** "l'intensité thématique" **est** "faible" **Alors** "le thème" **est** "peu probable"

"l'intensité thématique" représente la variable d'entrée de l'inférence floue,"faible" représente le terme linguistique associé à la variable d'entrée, "le thème" représente la variable de sortie de l'inférence floue et "peu probable" représente le terme linguistique associé à la variable de sortie.

L'ensemble des règles utilisées ici peut être représenté sous la forme d'une matrice comme sur la figure 5.9 où les entrées floues sont représentées en ligne (l'intensité thématique $I_{Drame}$) et en colonne (la richesse) par leurs symboles linguistiques. La variable linguistique de sortie **Drame** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que le synopsis traduise la thématique du Drame. Les valeurs prises par la variable de sortie (notée (2) sur la figure 5.6) sont représentées dans chacune des cellules de cette matrice. Si on prend la première cellule (en haut à gauche règle N°1) et la dernière cellule (en bas à droite règle N°25) de la matrice de la figure 5.9 les règles sont par exemple les suivantes :

| Numéro | Règle |
|---|---|
| 1 | **Si** ($I_{Drame}$ est TF) **ET** (richesse est TC) **Alors** (Drame est Faible) |
| 13 | **Si** ($I_{Drame}$ est M) **ET** (richesse est M) **Alors** (Drame est Moyen) |
| 14 | **Si** ($I_{Drame}$ est M) **ET** (richesse est L) **Alors** (Drame est Haut) |
| 25 | **Si** ($I_{Drame}$ est TH) **ET** (richesse est TL) **Alors** (Drame est Haut) |

| | Richesse | | | | |
|---|---|---|---|---|---|
| iDrame | TC | C | M | L | TL |
| TF | *F* | *F* | *F* | *F* | *F* |
| F | *F* | *F* | *F* | *F* | *H* |
| M | *F* | *M* | *M* | *H* | *H* |
| H | *M* | *H* | *H* | *H* | *H* |
| TH | *H* | *H* | *H* | *H* | *H* |

FIGURE 5.9 – Règles de combinaison entre l'intensité thématique du Drame et de la richesse du synopsis pour obtenir la mesure du Drame représentée par 3 symboles **F**aible, **M**oyen, **H**aut

Ces règles sont utilisées pour simuler dans le système flou le raisonnement des experts. Elles traduisent symboliquement la connaissance ou expertise du domaine et sont obtenues à partir du raisonnement d'un expert, exprimé dans le langage naturel.

#### 5.3.1.3 L'inférence floue

Un système flou est un système qui émule le raisonnement d'un expert à l'intérieur d'un domaine spécifique de connaissances. Ce raisonnement flou permet d'inférer la sortie à partir

des règles expertes et il s'effectue selon le principe de combinaison-projection dénommé "Zadeh's compositional rule of inference" ou « modus ponens généralisé » [Mauris *et al.*, 1996]. Son expression, dans le cas d'un système à deux entrées (ici l'intensité thématique et la richesse), représentée sous forme symbolique, est la suivante :

$$\forall Y \in L_Y, \mu_F(Y) = \perp_{(X_1,X_2)\in L_{X_1}\times L_{X_2}} \top_1\big(\top_2(\mu_{E_1}(X_1), \mu_{E_2}(X_2)), \mu_\Gamma(X_1, X_2, Y)\big) \quad (5.6)$$

où dans cette expression :

- $L_{X_1}$ et $L_{X_2}$ sont les univers linguistiques décrivant les entrées. Soit dans notre exemple :

$$L_{X_1} = L_{I_{Drame}} = \{\text{"Très Faible", "Faible", "Moyenne", "Haute", "Très Haute"}\}$$
$$L_{X_2} = L_{richesse} = \{\text{"Très Court", "Court", "Moyen", "Long", "Très Long"}\}$$

- $X_1$ et $X_2$ sont les variables linguistiques des entrées prenant leurs valeurs respectivement dans les univers linguistiques $L_{X_1}$ et $L_{X_2}$.

- $\mu_{E_1}(X_1)$ et $\mu_{E_2}(X_2)$ sont les degrés d'appartenance des valeurs numériques $x1$ et $x2$ aux variables linguistiques $X_1$ et $X_2$.

- $L_Y$ est l'univers linguistique décrivant la sortie. Soit dans notre exemple :

$$L_Y = L_{Drame} = \{\text{"Faible", "Moyen", "Haut"}\}$$

- $Y$ est la variable linguistique de sortie prenant ses valeurs dans l'univers linguistiques $L_Y$.

- $\mu_\Gamma(X_1, X_2, Y)$ représente les règles symboliques floues. Ce terme vaut 1 quand la règle liant $X_1$, $X_2$, $Y$ existe sinon il vaut zéro. Nous n'avons pas considéré de pondération des règles.

- $\top_1$ est l'opérateur de **combinaison** ou de *modus ponens généralisé*.

- $\top_2$ est le produit cartésien.

- $\perp$ est l'opérateur de **projection**.

L'étape de combinaison consiste à agréger de manière conjonctive à l'aide d'une t-norme le terme $\mu_\Gamma(X_1, X_2, Y)$ avec le résultat de l'agrégation des prémisses réalisées par l'opérateur $\top_2$. Cette combinaison va aboutir à un ensemble de poids sur l'appartenance de l'élément observé caractérisé par les mesures $x1$ et $x2$ au symbole $Y$. La projection de ces degrés est réalisée par l'opérateur de projection qui est un opérateur disjonctif de type t-conorme. Cette combinaison/projection est répétée pour toutes les classes de sortie. Le résultat final est donc un résultat flou composé de degrés d'appartenance affectés à chaque symbole de sortie. De plus, l'équation 5.6 de combinaison/projection peut être étendue à $n$ entrées puisque les opérateurs utilisés sont associatifs.

$$\top(x, \top(y, z)) = \top(\top(x, y), z)$$

Comme nous l'avons déjà dit, le choix des partitionnements des univers de discours $L(X)$ a été effectué dans le but de modéliser le raisonnement des experts, ce qui se traduit par une somme de l'ensemble des degrés égale à un (voir l'équation 5.5). Pour conserver cette cohérence entre les entrées et la sortie nous choisissons de conserver cette contrainte sur la sortie inférée. Il a été montré dans [Mauris *et al.*, 1996] que les opérateurs de combinaison et de projection avaient une influence sur cette contrainte. Finalement nous choisissons d'implémenter les opérateurs suivants :

$$\begin{aligned} \top(x,y) &= x * y \\ \bot(x,y) &= min(x+y,1) \end{aligned} \tag{5.7}$$

#### 5.3.1.4 Illustration

Nous présentons ici un exemple de raisonnement flou à partir du système de fusion que nous venons de décrire. Nous fixons les valeurs numériques d'entrée $I_{Drame} = 7.8$ pour l'intensité thématique et $richesse = 17.6$ pour la richesse du texte. La fuzzification de ces entrées permet à partir des fonctions d'appartenance (figure 5.7 et figure 5.8) d'obtenir la description floue de chacune des entrées :

$$F_{Drame}(7.8) = 0/\text{Très Faible} + 0.4/\text{Faible} + 0.6/\text{Moyen} + 0/\text{Haute} + 0/\text{Très Haute}$$
$$F_{richesse}(17.6) = 0/\text{Très Court} + 0/\text{Court} + 0.8/\text{Moyen} + 0.2/\text{Long} + 0/\text{Très Long}$$

On voit à partir de ces descriptions que la mesure de l'intensité thématique et de la richesse du synopsis sont décrites respectivement par deux variables linguistiques (Faible,Moyen) et (Moyen,Long). La fusion de ces informations symboliques est faite en utilisant le mécanisme d'inférence qui utilise les règles définies par expertise (voir la figure 5.9). A partir de ce jeu de symboles seules 4 règles sont actives (voir figure 5.10) :

| iDrame \ Richesse | TC | C | M | L | TL |
|---|---|---|---|---|---|
| TF | *F* | *F* | *F* | *F* | *F* |
| F | *F* | *F* | *F* | *F* | *H* |
| M | *F* | *M* | *M* | *H* | *H* |
| H | *M* | *H* | *H* | *H* | *H* |
| TH | *H* | *H* | *H* | *H* | *H* |

FIGURE 5.10 – Les quatre règles actives (en gras) lorsque $I_{Drame} = 7.8$ et $richesse = 17.6$

| Numéro | Règle |
|---|---|
| 8 | **Si** ($I_{Drame}$ est F) **ET** (richesse est M) **Alors** (Drame est Faible) |
| 9 | **Si** ($I_{Drame}$ est F **ET** (richesse est L) **Alors** (Drame est Faible) |
| 13 | **Si** ($I_{Drame}$ est M) **ET** (richesse est M) **Alors** (Drame est Moyen) |
| 14 | **Si** ($I_{Drame}$ est M) **ET** (richesse est L) **Alors** (Drame est Haut) |

Le résultat de l'inférence à partir de la règle de combinaison/projection (voir équation 5.6) et des opérateurs définis précédemment (voir équation 5.7) est :

$$F_{Drame}(I_{Drame} = 7.8, richesse = 17.6) = 0.4/\text{Faible} + 0.48/\text{Moyen} + 0.12/\text{Haute}$$

Ce résultat est présenté sous une forme floue traduisant une gradualité dans l'appartenance à chacun des concepts recherchés. Il peut être interprété de la façon suivante : le synopsis $S$ d'intensité thématique $I_{Drame}$ et de richesse $richesse$ appartient avec un degré de 0.4 au concept *« le synopsis à une possibilité* ***Faible*** *de traduire du Drame »*. Il appartient aussi, avec un degré de 0.48, au concept *« le synopsis à une possibilité* ***Moyenne*** *de traduire du Drame »* et enfin il appartient avec un degré de 0.12 au concept *« le synopsis à une possibilité* ***Haute*** *de traduire du Drame »*.

Cependant le but de cette fusion est d'être capable de statuer sur l'appartenance ou non du synopsis au thème du drame. Afin de transformer le résultat flou précédent sous une forme nette nous choisissons de défuzzifier ce résultat en recherchant le degré maximal présent en sortie. Ainsi la classe de sortie ayant le degré d'appartenance maximal sera attribuée au synopsis. Cette façon de procéder semble assez naturelle du fait de la propriété imposée sur la sortie (voir équation 5.5). Finalement un synopsis sera considéré comme appartenant au drame si sa classe de sortie est **Haut** ou **Moyen** et s'exprime par la relation suivante :

$$C_{out} = \arg\max_{L \in L_{out}}(\mu_{out}(L)) \tag{5.8}$$

$$Drame(S) = \begin{cases} 1 & \text{si } C_{out} \in \{\text{Moyen,Haut}\} \\ 0 & \text{sinon} \end{cases}$$

où le partitionnement de l'univers de discours de la sortie est $L_{out}(X) = \{\text{Faible,Moyen,Haut}\}$.

Nous décidons de classifier les 5804 synopsis de la base grâce à cette règle de classification. Nous comparons les résultats de classification avec le genre déclaré. Si le genre déclaré est le drame alors le classifieur à retrouvé le genre du film sinon il s'est trompé. Nous obtenons la matrice de confusion (voir tableau 5.1) où chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe déclarée (ou de référence).

| Déclaré \ Estimé | *NonDrame* | *Drame* |
|---|---|---|
| *NonDrame* | 2672 *(VN)* | 834 *(FP)* |
| *Drame* | 311 *(FN)* | 246 *(VP)* |

TABLE 5.1 – Matrice de confusion sur la prédiction du Drame. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

A partir de cette matrice de confusion nous calculons deux indicateurs qui sont la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP+FP} = \frac{246}{246+834} = 23\%$$
$$\text{Rappel} = \frac{VP}{VP+FN} = \frac{246}{246+311} = 44\%$$

Nous utilisons également le F-score (ou F-mesure) qui combine la précision et le rappel :

$$F_{score} = 2 * \frac{P * R}{P + R} = 2 * \frac{24 * 34}{24 + 34} = 30\%$$

On remarque que ces taux restent relativement faibles. Cependant l'utilisation et la fusion de l'information de richesse du synopsis a amélioré les résultats de rappel (augmentation de 30%) par rapport à l'approche basée uniquement sur l'intensité thématique (voir figure 4.25).

Finalement nous répétons ce principe de fusion d'information et de classification pour les couples d'entrée $\{I_{Policier}$, richesse du synopsis$\}$ et $\{I_{Humour}$, richesse du synopsis$\}$. Les définitions des partitions floues, les règles de fusion et les résultats de classification (F-score de 40% pour la thématique du "policier" et de 17 % pour la thématique de "l'humour"sont disponibles en annexe E.1.

#### 5.3.1.5 Fusion Texte/Texte

Nous disposons de trois informations symboliques qui traduisent la possibilité du synopsis à dégager une atmosphère liée au drame, au policier et à l'humour. Nous avons vu dans le chapitre précédent que les synopsis traitant de c'est deux dernières atmosphères utilisaient parfois un vocabulaire dramatique. Afin de diminuer l'incertitude sur la mesure du drame à partir des intensités thématiques il est intéressent de prendre en compte ces informations « parasites ». Puisque les informations sont déjà sous forme symboliques leur fusion nécessite de définir les règles de combinaison présentées sur la figure 5.11, où chaque tableau représente l'ensemble des combinaisons possibles entre les variables du "Policier" (en colonne) et de "l'Humour" (en ligne) et cela pour chacune des valeurs prises par la variable "Drame" (soit 3 valeurs et donc un tableau lorsque Drame=Faible, un tableau lorsque Drame=Moyen, un tableau lorsque Drame=Haut).

La fusion de ces informations symboliques est obtenue par le principe de combinaison/projection. La variable linguistique de sortie $Drame_{Texte}$ est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que le synopsis traduise du drame. Les valeurs prises par cette variable de sortie (notée (7) sur la figure 5.6) sont représentées dans chacune des cellules de cette matrice. On voit sur la figure 5.11 que plus le synopsis utilise un vocabulaire dramatique sans utiliser les vocabulaires de la thématique du policier et de l'humour plus la possibilité qu'il soit du drame (information de sortie) est élevée. Cette fusion exclusive permet de diminuer la certitude de l'information inférée. On voit ici toute la puissance de ce système de fusion car les règles de combinaison sont claires et interprétables (voir même triviales).

L'information obtenue après fusion ($Drame_{Texte}$ notée (7)) est utilisée pour classifier les films. Les résultats et leurs commentaires sont présentés dans §5.3.4.

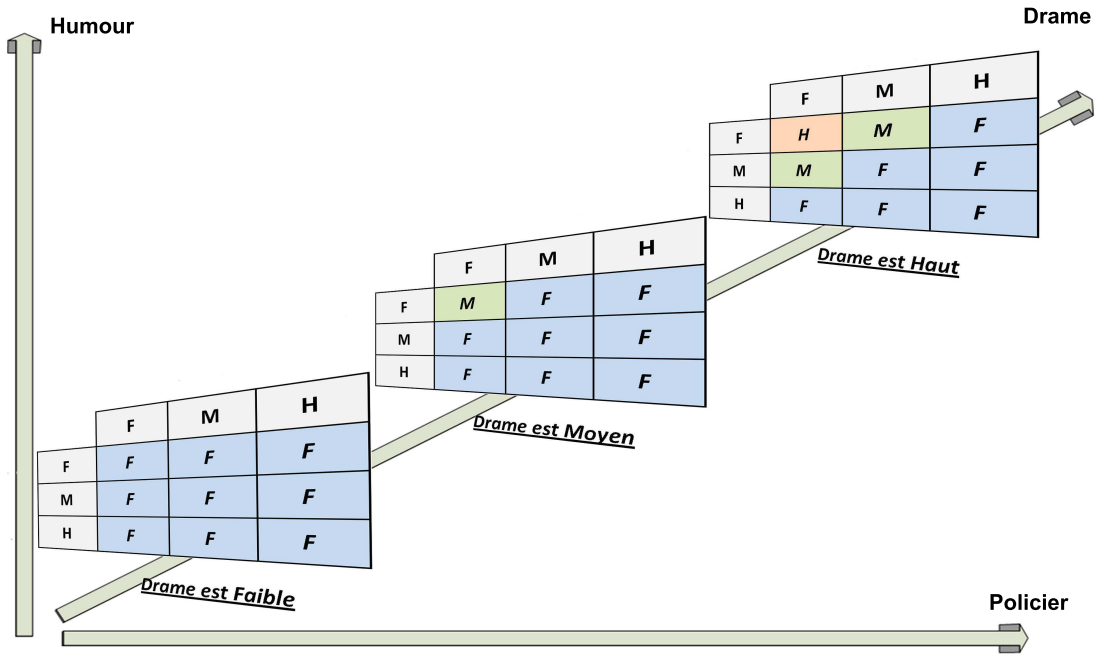



FIGURE 5.11 – Règles de combinaison pour obtenir la variable linguistique de sortie $Drame_{Texte}$ (information dans les cellules) entre les informations du Policier (en colonne) de l'Humour (en ligne) pour chacune des valeurs de la thématique du "Drame". Ces variables sont représentées par 3 symboles **F**aible, **M**oyen et **H**aut.

### 5.3.2 Fusion des indicateurs image

Nous avons présenté dans le chapitre consacré aux images un certain nombre de descripteurs images liés aux caractéristiques colorimétriques de la séquence vidéo et à son activité. Nous présentons ici le système de fusion qui permet de passer de ces descripteurs numériques à une information caractérisant l'atmosphère dramatique dégagée par la séquence d'images. Pour cela nous utilisons quatre informations issues de l'analyse des images :

**Le ratio de Couleurs Foncées** est utilisé pour caractériser la proportion de couleurs sombres (voir le §3.1.2.1 et le chapitre 7 de [Ionescu, 2007] pour le détail de son calcul). Cette information est utile pour définir différents concepts comme la **froideur** ou la **monotonie**.

**Le ratio de Couleurs Chaudes** est utilisé pour caractériser la proportion de couleurs chaudes (voir le §3.1.2.1 et le chapitre 7 de [Ionescu, 2007] pour le détail de son calcul). Cette information permet de donner la proportion de couleurs chaudes présentes dans la séquence et donc, par opposition, fournit l'absence de couleurs chaudes ce qui permet définir le concept de **froideur**.

**Le ration de Variation des Couleurs** est utilisé pour caractériser la richesse de la palette couleur utilisée pour composer le film (voir chapitre 7 de [Ionescu, 2007] pour le détail

de son calcul). Cette information est utile pour définir le concept de **monotonie**.

**L'Activité globale** est utilisée pour caractériser la fréquence du changement de contenu dans les images. Cette information est utile pour définir le concept de **monotonie**.

Nous fusionnons ces différentes sources d'informations numériques issues de l'analyse des images pour obtenir des descripteurs liés au drame. Ces nouvelles informations caractérisent les concepts artistiques utilisés pour construire une atmosphère noire, inquiétante dans les films d'animation. Ces différents concepts sont issus d'une expertise symbolique formulée dans un langage naturel et sont présentées ci-après :

**La froideur** est un concept qui caractérise les images où il y a une forte dominance de couleurs *froides* et *sombres*. Cette information est obtenue à partir d'un système flou (notée (4) sur la figure 5.6) dont le principe est identique à celui vu précédemment. Bien sûr, les partitions floues et les règles de combinaison sont adaptées au contexte et donc aux informations images. Ces éléments sont disponibles dans l'annexe E.1.3.

**La monotonie** est un concept qui caractérise les séquences d'images où il y a une forte dominance de couleurs *foncées* et où l'activité est *faible*. Cette information qui traduit une atmosphère lente et noire est obtenue à partir d'un système flou (notée (5) sur la figure 5.6) dont le principe est identique à celui vu précédemment. Les partitions floues et les règles de combinaison sont adaptées aux informations images et sont disponibles dans l'annexe E.1.4.

**L'uniformité** est un concept qui caractérise les séquences d'images où il y a une forte dominance de couleurs *froides* et où il y a une faible *variété* dans la palette couleur utilisée pour composer le film. Cette information traduit une atmosphère pauvre et froide en termes de couleur et est obtenue à partir d'un système flou (notée (6) sur la figure 5.6) dont le principe est identique à ceux vus précédemment. Les partitions floues et les règles de combinaisons sont adaptées aux informations images et sont disponibles dans l'annexe E.1.5.

Ces concepts définis à partir des valeurs numériques images sont liés au drame et permettent la caractérisation de l'atmosphère du film. Ainsi pour obtenir une information moins incertaine nous proposons de fusionner ces concepts.

#### 5.3.2.1 Fusion Image/Image

Nous disposons de trois informations symboliques traduisant la possibilité de la séquence d'images à dégager une atmosphère liée à la froideur, à la monotonie et à l'uniformité, atmosphères liées à une atmosphère dramatique. Ainsi il est intéressant de les utiliser conjointement pour définir la possibilité que les images dégagent une atmosphère dramatique. Puisque ces informations sont déjà sous forme symbolique leur fusion nécessite de définir des règles de combinaison. Ces règles de combinaison sont présentées sur la figure 5.12 où chaque tableau

représente l'ensemble des combinaisons possibles entre les variables de l'"Uniformité" (en colonne) et de la "Monotonie" (en ligne) et cela pour chacune des valeurs prises par la variable de la "Froideur" (soit 3 valeurs et donc un tableau lorsque Froideur=Faible, un tableau lorsque Froideur=Moyen, un tableau lorsque Froideur=Haut).

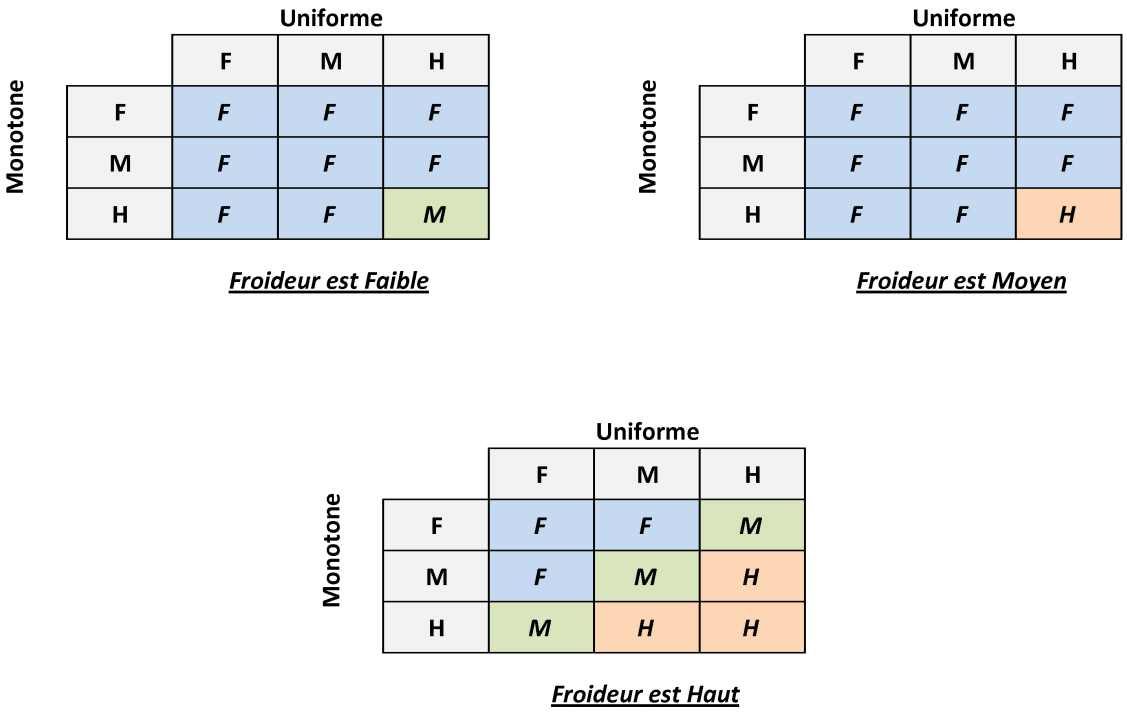



FIGURE 5.12 – Règles de combinaison pour obtenir la variable linguistique de sortie $Drame_{Image}$ (information dans les cellules) entre les informations de l'Uniformité (en colonne) et de la Monotonie (en ligne) pour chacune des valeurs de la "Froideur". Ces variables sont représentées par 3 symboles **F**aible, **M**oyen et **H**aut.

La fusion de ces informations symboliques est obtenue par le principe de combinaison/projection. La variable linguistique de sortie $Drame_{Image}$ est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que les images traduisent du drame. Les valeurs prises par cette variable de sortie (notée (8) sur la figure 5.6) sont représentées dans chacune des cellules de cette matrice. Ainsi on voit sur la figure 5.12 qu'un film a une possibilité avec une appartenance de type moyenne ou élevée d'être dramatique si il a au moins deux des trois concepts précédents dont la possibilité est élevée. Naturellement, sa possibilité devient élevée si le film contient les trois concepts en même temps avec une possibilité élevée. Finalement, l'information de froideur qui est la plus caractéristique du drame (en termes de certitude) a une importance supérieure par rapport aux deux autres concepts. Cette remarque est visible sur la troisième matrice où la possibilité de la froideur est élevée. En effet, on voit que la possibilité du drame est de type Moyen même si les informations de monotonie et d'uniformité sont moyennes.

L'information obtenue après fusion ($Drame_{Image}$ notée (8)) est utilisée pour classifier les films. Les résultats et leurs commentaires sont présentés au §5.3.4.

### 5.3.3 Fusion du texte et de l'image

Nous disposons à présent de deux informations symboliques caractérisant la possibilité du film d'appartenir au genre du drame. Il reste donc à combiner ces informations issues du texte et de l'image pour **diminuer l'incertitude** de chacune de ces informations prises indépendamment. Cette fusion d'information est assurée par le dernier étage de notre système de fusion présenté sur la figure 5.6 et correspond à la sortie *(9)*.

Pour réaliser cette combinaison nous appliquons le principe de combinaison/projection qui a déjà été défini (voir équation 5.6). Ainsi nous définissons la variable linguistique de sortie $Drame_{Fusion}$ qui constitue le résultat final de fusion et qui est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que le film traite d'un sujet dramatique. Les valeurs prises par cette variable de sortie sont disponibles dans chacune des cellules de la matrice (voir figure 5.13) qui représente la connaissance des experts.

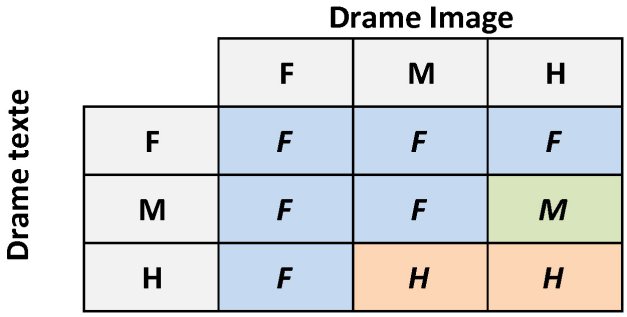



FIGURE 5.13 – Règles de combinaison entre les informations textuelle et image représentées par 3 symboles **F**aible, **M**oyen, **H**aut

On voit sur la figure 5.13 qu'un film a une possibilité élevée d'être dramatique si chacune de ses deux sources ont une possibilité élevée d'être du drame (informations concordantes). Or on l'a vu, les informations textuelles sont des informations de plus haut niveau sémantique que les informations images. En réalité, l'intensité thématique permet de mesurer quasi directement des concepts qui sont proches de l'atmosphère recherchée alors que les paramètres de couleurs ou d'activité ne mesurent que les traces d'une « norme » de création liée à cette atmosphère noire, inquiétante. Par conséquent, même si ces deux informations mesurent les traces d'une volonté artistique de plonger le spectateur dans l'atmosphère dramatique, l'incertitude concernant l'information textuelle est moins importante que celle concernant les images. Finalement, l'information du drame à partir du texte a une importance supérieure dans le résultat final par rapport à l'information du drame issue des images. Cette confiance accordée à l'information textuelle est visible dans les règles de combinaison suivantes :

| Numéro | Règle |
|---|---|
| 6 | **Si** ($Drame_{Texte}$ est H) **ET** ($Drame_{Image}$ est M) **Alors** ($Drame_{Fusion}$ est H) |
| 8 | **Si** ($Drame_{Texte}$ est M) **ET** ($Drame_{Image}$ est H) **Alors** ($Drame_{Fusion}$ est M) |

### 5.3.4 Test et résultats

Pour vérifier le pouvoir discriminant des informations obtenues tout au long du processus de fusion nous décidons de classifier, à partir des informations inférées, les films de la base dont nous disposons. La prise de décision quant à l'appartenance pour un film au genre dramatique est représentée par l'équation 5.8. Ainsi, si un film est caractérisé par l'information $I_{out}$ alors ce film sera considéré comme appartenant au drame si la classe de sortie de l'information $I_{out}$ est **Haut** ou **Moyen**.

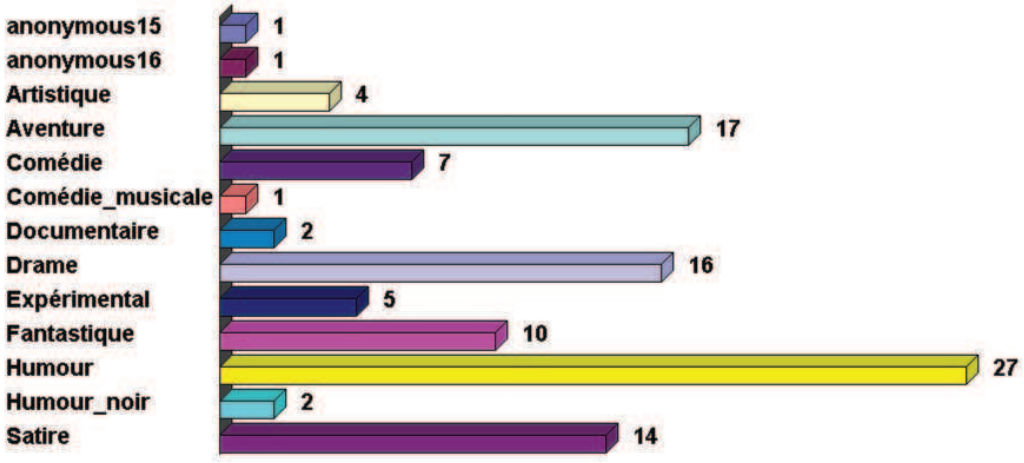



FIGURE 5.14 – Répartition des 107 films suivant le ou les genres déclarés

Nos tests sont effectués sur une base de 107 films d'animation (107 synopsis) et la figure 5.14 représente la répartition des films suivant le genre déclaré (voir également l'annexe E.2 pour quelques caractéristiques de cette base). Cet échantillon de la base de CITIA reste limité car malheureusement nous ne disposons pas des droits d'auteur et d'exploitation des films sur la totalité de la base de CITIA.

La figure 5.15 présente les résultats des différents tests de classification où les informations qui ont permis la classification des films sont représentées en colonne et où les mesures de précision et de rappel sont représentées en ligne. Les numéros des informations correspondent aux numéros de la figure 5.6.

| *Règle de décision: C(out) = {Moyen,Haut}* | Intensité thématique du drame seuillée à 8.3 | Drame (2) | Fusion Texte/texte (7) | Froideur (4) | Monotone (5) | Uniforme (6) | Fusion Image/Image (8) | Fusion Texte/Image (9) |
|---|---|---|---|---|---|---|---|---|
| Précision | 41% | 41% | 43% | 28% | 20% | 28% | 39% | 78% |
| Rappel | 81% | 88% | 81% | 81% | 69% | 85% | 56% | 44% |
| F-score | 54% | 56% | 56% | 42% | 31% | 42% | 46% | 56% |
| | | Texte | | | Image | | | |

FIGURE 5.15 – Résultats de classification de la prédiction du Drame en termes de Précision, Rappel et F-Score (représentes en ligne), en fonction de l'information utilisée (représentée en colonne) pour les valeurs symboliques "Moyen" et "Haut"

On remarque que la fusion de l'intensité thématique du drame et de la richesse du synopsis permet d'améliorer le rappel et de le faire passer de 81% (dans le cas d'un classification à partir de l'intensité thématique voir la figure 4.6) à 88%. La précision quant à elle n'est pratiquement pas modifiée ce qui est normal car l'utilisation de la richesse permet en quelque sorte d'ajuster le seuil de l'intensité thématique. Cela a pour conséquence de retrouver beaucoup plus de synopsis dramatiques qu'avec un seuil fixe (comme c'est le cas avec la méthode utilisée dans la première colonne). L'utilisation et la fusion des informations liées aux thèmes du Policier et de l'Humour permet d'améliorer la précision de la classification (de 41% à 43%). En effet, ces informations permettent de ne considérer comme dramatique que les textes qui ont exclusivement un vocabulaire lié au drame. Cela fait chuter le rappel (de 88% à 81%) car comme on l'a vu sur l'analyse des textes certains films dramatiques ont des synopsis qui mélangent les vocabulaires. Finalement l'information textuelle permet de retrouver avec un bon taux de rappel (de plus de 80%) les films dramatiques, par contre elle est très incertaine puisque la précision n'est que de 43%.

Les remarques sont identiques dans le cas des descripteurs images. En effet, les concepts comme la froideur ou l'uniformité permettent de retrouver avec un bon taux de rappel les films dramatiques ce qui indique que ces films d'animation suivent bien le principe artistique qui veut que l'atmosphère dégagée par un film est liée aux couleurs choisies pour le composer. Cependant ce principe est également utilisé pour d'autres catégories de films comme le montre la mesure de précision. Finalement, la fusion de ces trois sources permet d'améliorer significativement la précision de la classification (de 25% en moyenne à 39%) et donc de diminuer l'incertitude de l'information image.

Finalement la fusion du texte et de l'image qui doit permettre de diminuer l'incertitude de l'information, permet de classifier les films dramatiques avec une bonne précision (près de 80%). La fusion de ces deux sources complémentaires (complémentarité confirmée par la mesure $Q_{i,k} = -0.2$ de dépendance entre les classifieurs image et texte voir §5.1.3) permet d'améliorer significativement la précision de la classification (de 41% en moyenne à 78%) et donc de diminuer l'incertitude de l'information image et texte. Cependant le rappel n'est pas très bon ce qui indique que tous les films ne sont pas retrouvés. De plus, lorsque l'on s'intéresse aux films abusivement considérés comme du drame (les Faux Positifs) on trouve les films suivants :

**La bouche cousue (1998)** : Genre déclaré : **Humour**
*« Un personnage, au regard **triste** et **perdu**, monte dans le bus avec une pizza dans les mains. Il est presque assis lorsque le chauffeur freine **brutalement**. ».*

**Vent (1964)** : Genre déclaré : **Aventure**
*« Un homme se **bat** contre une **tempête**. Lorsque, soudain, il rencontre une fillette, nous découvrons que quelqu'un contrôle le vent. ».*

Ces films sont tous les deux composés de couleurs froides et sombres impliquant une information $Drame_{Image}$ issue de l'image avec une possibilité élevée. De plus, ils utilisent tous les deux des termes à connotation dramatique ce qui implique une information $Drame_{Texte}$ issue

du texte avec une possibilité moyenne. Finalement la fusion de ces informations implique que ces films ont une possibilité moyenne d'être des films dramatiques (voir la figure 5.13) et sont classés comme dramatiques par la règle de classification.

Pour diminuer le nombre de Faux Positifs, nous décidons de changer la règle de prise de décision quant à l'appartenance d'un film au genre dramatique. Ainsi si un film est caractérisé par l'information $I_{out}$ alors ce film sera considéré comme appartenant au drame si la classe de sortie de l'information $I_{out}$ est **Haut**.

Les tests sont effectués sur la même base de 107 films d'animation et la figure 5.16 présente les résultats dans les mêmes conditions expérimentales que précédemment.

| *Règle de décision: C(out) = {Haut}* | Intensité thématique du drame seuillée à 8.3 | Drame (2) | Fusion Texte/texte (7) | Froideur (4) | Monotone (5) | Uniforme (6) | Fusion Image/Image (8) | Fusion Texte/Image (9) |
|---|---|---|---|---|---|---|---|---|
| Précision | 41% | 50% | 50% | 34% | 26% | 33% | 43% | 100% |
| Rappel | 81% | 50% | 50% | 63% | 56% | 63% | 56% | 25% |
| F-score | 54% | 50% | 50% | 44% | 36% | 43% | 49% | 40% |
| | Texte | | | Image | | | | |

FIGURE 5.16 – Résultats de classification de la prédiction du Drame en termes de Précision, Rappel et F-Score (représentés en ligne), en fonction de l'information utilisée (représentée en colonne) pour la valeur symbolique "Haut"

On voit que la fusion du texte et de l'image permet d'avoir une excellente précision (100%), mais le rappel a chuté fortement (25%). Cela veut dire que l'information en sortie du système de fusion a une très bonne certitude mais cette information ne permet pas de retrouver tous les films du genre dramatique. Cela veut dire également que les sources d'informations utilisées ne permettent pas de discriminer l'ensemble des films dramatiques. Cela pose la question de savoir si il est finalement possible de déterminer le genre du film à partir de mesures réalisées sur des films aussi variés et aux contenus complexes en utilisant un système de fusion par raisonnement et expertise. Pour répondre à cette question nous comparons notre méthode de fusion avec une approche par classification automatique supervisée (méthode sans expertise). Deux algorithmes sont testés :

- Les réseaux de neurones de type "Multi-Layer Perceptron (MLP)" sont couramment utilisés dans des problèmes de classification supervisée dans [Caicedo *et al.*, 2008] ou bien encore dans [Tsai et Wu, 2008]. Le MLP est une extension multicouche du perceptron (réseau à une couche, assez limité). Il utilise un algorithme d'apprentissage très répandu basé sur la mesure de l'erreur quadratique moyenne baptisé rétropropagation du gradient d'erreur. Nous avons utilisé l'algorithme disponible dans le logiciel de classification Weka [Witten *et al.*, 1999].

- Les SVM constituent également une famille de classifieurs couramment rencontrés dans différents travaux comme dans [Tong et Chang, 2001, Caicedo *et al.*, 2008]. Ces techniques ont été décrites dans [Vapnik, 1996] et consistent à délimiter par la frontière la plus large possible les différentes catégories des échantillons de l'espace vectoriel du

corpus d'apprentissage. Les vecteurs supports constituent les éléments délimitant cette frontière. Nous avons utilisé l'implémentation du type Sequential Minimal Optimization (SMO) [Platt, 1999] avec comme fonction noyau le "cubic polynomial kernel" ou polynôme de degré trois disponible sous Weka [Witten *et al.*, 1999].

La classification du genre dramatique est réalisée à partir des informations numériques texte et image présentées précédemment. Nous avons opté pour un test par validation croisée de paramètre trois[3].

| **Méthode⟶** | **Fusion Floue** | **SVM** | **MLP** |
|---|---|---|---|
| **Précision** | 78% | 47% | 44% |
| **Rappel** | 44% | 68% | 44% |
| **F-Score** | 56% | 56% | 44% |

TABLE 5.2 – Comparaison de méthodes de fusion.

On voit à partir de ce test comparatif (tableau 5.2) que notre méthode par expertise et la méthode par SVM donnent des résultats identiques (F-Score de 56%) bien meilleurs que les résultats du réseau de neurones (F-Score de 44%). La mesure de rappel de notre système n'est pas catastrophique puisqu'elle est identique à la méthode du MLP. Seule la méthode de type SVM permet d'obtenir une mesure de rappel relativement bonne (68%). Cependant, notre méthode de fusion permet d'obtenir une précision bien meilleure que celle obtenue par l'approche de type SVM. Cela veut dire que l'information en sortie de notre système de fusion est moins incertaine que celle fournie par le classifieur.Cependant même si cette méthode retrouve 11 films sur 16 elle ne permet pas de retrouver l'ensemble des films.

En conclusion, il paraît difficile de déterminer le genre de tous les films à partir des seules sources d'informations retenues dans ce travail. Soit il est nécessaire d'ajouter de nouvelles informations (modalité son par exemple), soit la variété des films et des contenus constitue une frontière difficilement franchissable sans un processus cognitif complexe. Les résultats de précision de notre méthode sont encourageants car il démontre que pour la moitié des films dont nous disposons, les informations de couleur, d'activité et d'intensité thématique permettent avec une bonne certitude de retrouver les films dramatiques. Enfin, notre système a l'avantage d'être complètement compréhensible et permet d'expliquer les résultats car les règles de combinaisons sont facilement explicitables, ce qui est très important pour valider notre approche auprès des professionnels de l'animation. De plus, même si l'expertise n'est pas facile à formaliser la constitution du système de fusion à partir de connaissances *a priori* a un coût plus faible (en termes de temps passé par exemple ) que la constitution bien souvent fastidieuse d'une base d'apprentissage numérique, exhaustive, etc. utilisée pour l'apprentissage de classifieur comme c'est souvent le cas dans le domaine de l'image et la vidéo (par exemple dans TRECVid où des centaines d'heures de films sont annotés manuellement).

3. La base est découpée en trois parties égales, 2 parties de la base sont utilisées pour l'apprentissage, la partie restante (1/3) est utilisée pour tester le classifieur appris. Cette opération est répétée 3 fois pour que chaque partie soit utilisée en apprentissage et en test puis les résultats sont moyennés

## 5.4 Caractérisation locale des films appliquée à l'activité

Notre second objectif est de caractériser d'un point de vue local[4] le film à travers une information qui est l'action. En effet nous désirons dans cette section exploiter conjointement le scénario actanciel et les séquences d'images. Ce choix est motivé par le fait que le scénario actanciel permet le plus souvent de décrire et nommer l'action, les actants et la scène présentés dans la sous-séquence vidéo. Ces informations sont d'un très haut niveau sémantique par rapport à l'information qu'il est possible d'extraire des images (voir le chapitre consacré aux images). Cette ressource textuelle est donc très intéressante à exploiter pour augmenter la sémantique des informations issues des images. Malheureusement cette ressource est complètement désynchronisée par rapport à la vidéo. En effet, les descriptions du scénario actanciel mais plus généralement les descriptions faites par le synopsis ne sont pas repérées sur l'axe temporel de la vidéo. Par conséquent, on sait que l'action existe probablement dans les images mais on ne sait pas à quel moment. Une première étape consiste donc à aligner ces deux sources (image et scénario actanciel) afin de retrouver la sous-séquence vidéo décrite par le texte. Peu de travaux sont proposés dans cette optique car les ressources textuelles habituellement utilisées sont soit les textes incrustés dans l'image ou alors des méta données comme les sous titrages [Marszalek *et al.*, 2009] qui sont des informations parfaitement synchronisées avec le support vidéo.

L'idée développée dans nos travaux est de réaliser cette synchronisation à partir de l'**action**. En effet, cette information fondamentale du scénario actanciel peut être liée aux informations d'activité extraites de la séquence d'images. Ainsi, si nous arrivons à retrouver dans la séquence vidéo le passage de l'action décrite par le scénario actanciel, alors nous pourrons probablement retrouver et nommer dans l'image les autres éléments du scénario actanciel (sous réserve d'avoir les détecteurs image adéquates comme la détection de personnage). La difficulté majeure pour l'instant est d'arriver à faire le lien entre la mesure de l'activité et l'action décrite dans le scénario actanciel. Plus précisément, il serait intéressant de savoir si il existe un lien entre l'activité mesurée localement dans la séquence d'images et les verbes utilisés pour décrire le ou les passages du film. Intuitivement on imagine mal qu'un synopsis puisse décrire l'action d'une sous-séquence et que l'activité dans cette sous-séquence soit moins importante que dans le reste du film. Autrement dit, si il existe un lien entre les termes d'action et l'activité mesurée dans ces ou cette sous-séquence(s), alors cette mesure dans ce(s) passage(s) doit probablement être supérieure au reste du film, ce qui se traduit par une mesure élevée du rythme.

Pour faire ce lien entre les termes et l'activité locale mesurée dans le film, nous avons exploré deux pistes afin de répondre aux questions suivantes :

- *l'action mentionnée dans les synopsis correspond-elle au rythme de la séquence ?*
- *une action mentionnée dans un synopsis correspond-elle à une action* ***locale*** *(passage d'activité) mesurée dans la séquence ?*

Ces deux approches vont être traitées dans les sections suivantes

4. c'est-à-dire situé sur l'axe temporel de la vidéo

### 5.4.1 Quels liens établir entre le texte et les images ?

Nous cherchons à savoir si il existe une relation entre le rythme (ou changement d'activités) mesuré par les images et les termes employés dans le synopsis notamment à travers les verbes extraits du scénario actanciel. Dans une première phase nous cherchons à savoir quels sont les verbes dans les synopsis qui traduisent un rythme identifiable dans les séquences d'animation. L'idée est d'obtenir automatiquement les relations qu'il pourrait y avoir entre l'activité intrinsèque des termes (par exemple **"chanter"** a une activité intrinsèque plus faible que **"courir"**) et le rythme qu'ils entrainent sur l'ensemble de la séquence vidéo.

La première étape consiste à extraire les verbes conjugués de chacun des synopsis. Pour cela nous utilisons l'analyseur grammatical LG et son module statistique présentés dans le chapitre sur l'analyse des textes. Notre analyse porte sur les synopsis anglais de la base des 107 films d'animation dont nous disposons et permet l'extraction d'un peu plus de 250 verbes. De plus, comme de nombreux verbes renvoient à un même concept (relation d'hyperonymie) il est nécessaire de prendre en compte ces concepts globaux afin de maximiser les chances de trouver des liens significatifs entre le texte et l'image. Ainsi, pour chacun des verbes retrouvés dans le synopsis nous extrayons grâce à WordNet l'ensemble de ses hypernymes (voir figure 4.23). L'ensemble des verbes et de leurs hypernymes constituent une base d'un peu plus de 570 termes.

La deuxième étape consiste à constituer la base de tests. Pour chacun des 570 termes nous construisons un ensemble d'exemples dans lequel chaque terme (verbe du synopsis ou hypernyme) est associé à la mesure du rythme (sous forme symbolique) issue de l'analyse des images (voir chapitre sur l'analyse des images §3.2.4). Lorsque le synopsis contient un verbe (terme) ayant plusieurs hypernymes, on constitue autant d'exemples que de termes, les attributs (numéro du film, rythme) restant identiques (voir la figure 5.17).

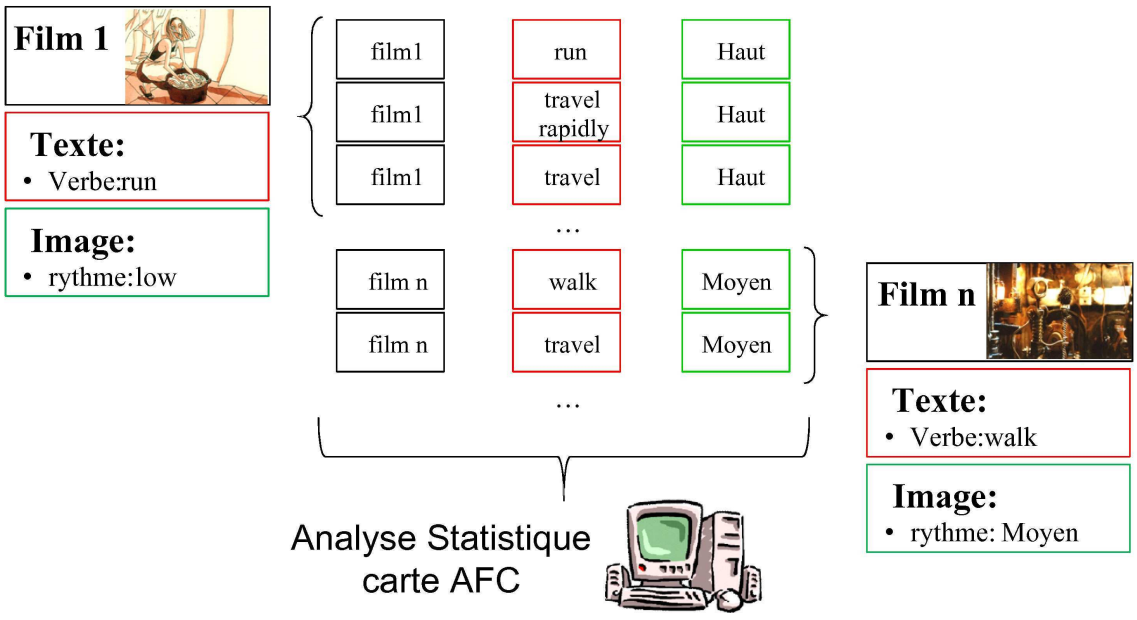



FIGURE 5.17 – Principe de constitution des exemples

La troisième étape consiste en la phase d'analyse statistique. Cependant parmi les 570 exemples il y a beaucoup d'hapax (terme dont l'occurrence est égale à un, verbes utilisés une seule fois) il est donc nécessaire de réduire l'espace de tests pour ne conserver que les termes les plus fréquents afin d'augmenter la pertinence des liens statistiques. Nous choisissons de ne conserver que les exemples dont l'occurrence du terme est supérieure ou égale à 10. Par conséquent, seuls les termes (hypernymes) suivants sont conservés :

| **Terme** | ***"act"*** | ***"be"*** | ***"change"*** | ***"make"*** | ***perceive"*** | ***"travel"*** |
|---|---|---|---|---|---|---|
| **Occurrence** | 26 | 50 | 12 | 25 | 12 | 35 |

Ils forment ainsi une base de test de 160 exemples. Cet ensemble d'exemples est analysé via une AFC afin de montrer les associations statistiques entre les verbes et le rythme symbolique (voir le chapitre sur l'analyse des textes §4.2.1.2). Cette analyse permet de déterminer s'il existe un lien privilégié ou non entre 2 termes par comparaison avec le cas d'indépendance qu'on aurait obtenu si les effectifs étaient répartis proportionnellement et indépendamment. Cette analyse est simple et bien connue, et fournit un cadre statistique (test du $Chi^2$) pour illustrer et valider les associations.

On peut donner une représentation plus visuelle des écarts à l'indépendance par l'utilisation d'une carte d'analyse factorielle des correspondances. Elle consiste à tracer une carte à partir des résultats de l'AFC en disposant les modalités en fonction des écarts à la situation d'indépendance. Par défaut, chaque modalité est représentée par un pavé de surface proportionnelle à son effectif. Leurs positions les unes par rapport aux autres permettent d'illustrer les propensions qu'ont les éléments à être associés.

La carte AFC de la figure 5.18 calculée sur les 160 exemples permet, à travers l'agencement des modalités et des constellations, d'identifier les associations entre les termes. Plus un verbe à tendance à être associé à un rythme symbolique plus ils seront proches. On remarque à partir de ces associations que les termes comme « travel » [5] traduisant du mouvement, « make » traduisant une action et « change » [6], traduisant une modification correspondent dans l'ensemble à un rythme qui est compris entre moyen et élevé. De plus, les verbes de perception comme « perceive » ou d'état comme « be » ont tendance à être associés à un rythme faible. Ceci est très intéressant car cela veut dire qu'un film dont le synopsis décrit une ou plusieurs action(s) est généralement associé à une mesure de rythme élevé. Les résultats de cette analyse permettent de mettre en lumière les liens qu'il y a entre notre mesure d'activité et l'utilisation de verbes d'action dans le synopsis. Plus généralement cela valide notre intuition : les images portent la trace des descriptions textuelles faites dans le synopsis. Cela traduit également que les actions décrites dans le synopsis se retrouvent probablement dans les images où les changements de contenu sont importants par rapport au reste de la séquence. L'étape suivante consiste à vérifier cette hypothèse et à caractériser de façon locale ce lien entre le synopsis et la mesure d'activité.

5. Synonymes extraits de WordNet : *travel, go, move, locomote – (change location ; move, travel, or proceed, also metaphorically)*

6. Synonymes extraits de WordNet : *change, alter, modify – (cause to change ; make different ; cause a transformation)*

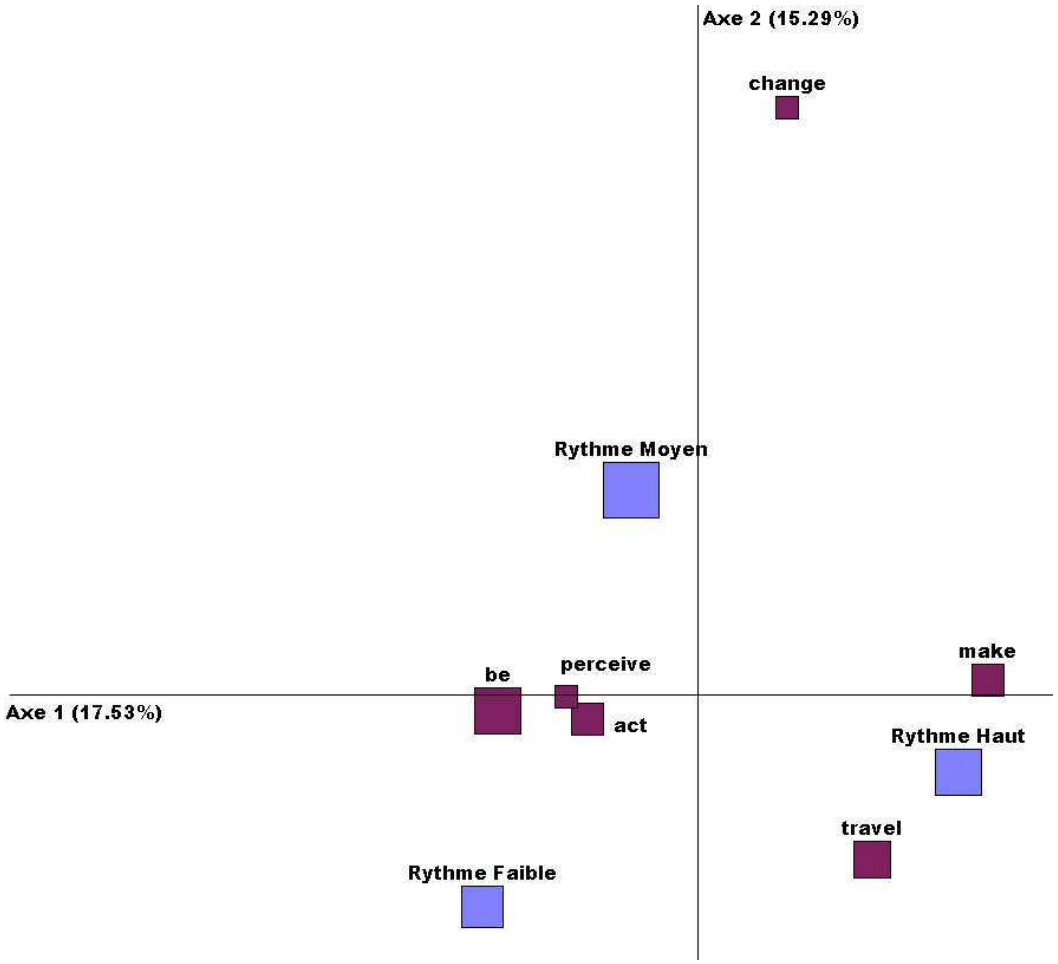



FIGURE 5.18 – Carte de l'Analyse factorielle des correspondances multiples pour la recherche d'associations entre le rythme symbolique et les verbes utilisés pour décrire le film.

### 5.4.2 Caractérisation de l'activité locale

Nous venons de voir de façon globale que les synopsis qui emploient des verbes de mouvement font probablement référence à un (des) passage(s) du film où il y a une activité élevée. Dans cette deuxième étude, nous avons étudié le synchronisme des termes utilisés dans le synopsis et les plages de la séquence d'animation correspondant à une activité locale élevée.

Pour vérifier et quantifier cette synchronisation (ou alignement) entre les deux sources d'information que sont les images (au travers des segments d'action locale) et le texte (au travers du scénario actanciel), nous proposons d'évaluer pour chacun des films le recouvrement temporel de ces informations. Cette étude effectuée manuellement est réalisée sur une vingtaine de films.

Dans un premier temps, nous retenons dans le synopsis anglais de chaque film les verbes d'action du scénario actanciel, c'est-à-dire les verbes qui ont comme concept hypernyme « travel » (voir la figure 5.19).

Dans un deuxième temps, nous repérons manuellement pour chaque film la ou les sous-séquence(s) où l'action décrite par le scénario actanciel est visible dans les images (ceci consti-

**(a) Le moine et le poisson** : *A monk finds a fish in the water reservoir of his monastery. He tries to catch it using all kinds of means and, as the film goes on, this becomes increasingly symbolic.*

**(b) L'homme aux bras ballants** : *In a sleepy town under a moonless sky, a character with enormous arms is walking. Preceded by his shadow, he makes his way to an arena in order to achieve a ritual..*

**(c) Tamer of Wild Horses** : *Will the man manage to tame the beast of iron and fire ? Yes, but only if it is without violence. Understood and loved, she takes the man to outer space.*

**(d) The flying man** : *A man is flying on the spot. Another man comes and tries to do the same but can't.*

FIGURE 5.19 – Synopsis anglais de quelques films d'animation. Les verbes de mouvement sont soulignés.

tue notre vérité terrain). Par exemple dans le film *Le moine et le poisson* nous cherchons manuellement la ou les sous-séquence(s) d'images où le moine tente d'attraper le poisson. Nous mettons ensuite en regard ces sous-séquences avec la mesure locale d'activité définie au paragraphe 3.2.4.2 dans le chapitre consacré aux images. Cette mise en correspondance est présentée pour quelques exemples sur la figure 5.20).

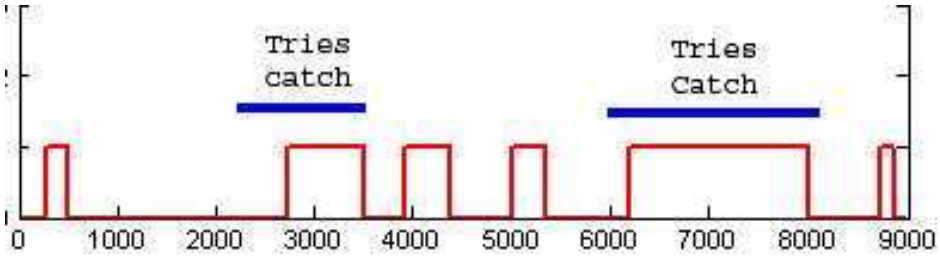
Sur l'ensemble des films analysés nous avons remarqué qu'il y a généralement adéquation entre la sous-séquence où l'activité locale est élevée et la sous-séquence où l'action est décrite par le(s) verbe(s) de mouvement. Les cas où il n'y a pas de synchronisation entre le descripteur image et l'action mentionnée correspondent à des synopsis ne faisant tout simplement pas référence à une partie de la séquence (synopsis généraliste).

Afin de mesurer les performances de cette synchronisation nous décidons de nommer automatiquement l'action mise en scène dans chacune des sous-séquences où l'activité est élevée par le ou les termes de mouvement retrouvé(s) dans le synopsis (voir la figure 5.21). Cette classification naïve des sous-séquences est ensuite comparée à la vérité terrain annotée manuellement (voir les segments bleus sur la figure 5.21). Si la sous-séquence d'activité locale est incluse dans la vérité terrain et si le terme la désignant est le même que celui de la vérité terrain alors cette sous-séquence est considérée comme correctement annotée. Ainsi nous obtenons les résultats présentés dans le tableau 5.3

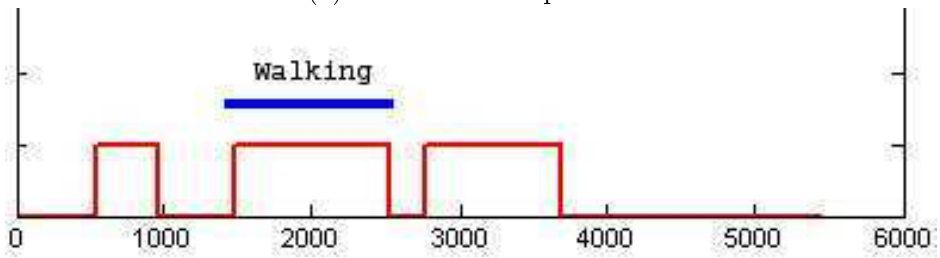
| | |
|---|---|
| **Précision** | 40% |
| **Rappel** | 90% |
| **F-Score** | 56% |

TABLE 5.3 – Résultats de l'alignement automatique de l'image et du texte à partir de l'activité locale.

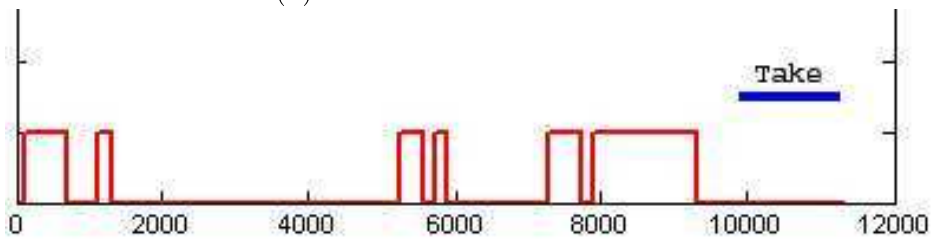
Nous voyons au travers de la mesure de Rappel du tableau 5.3 que l'on retrouve grâce



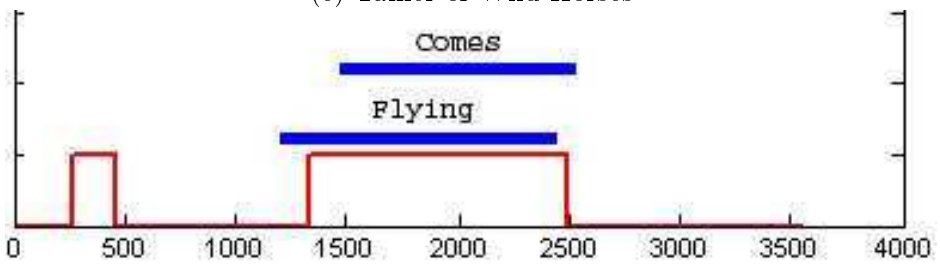

(a) Le moine et le poisson




(b) L'homme aux bras ballants



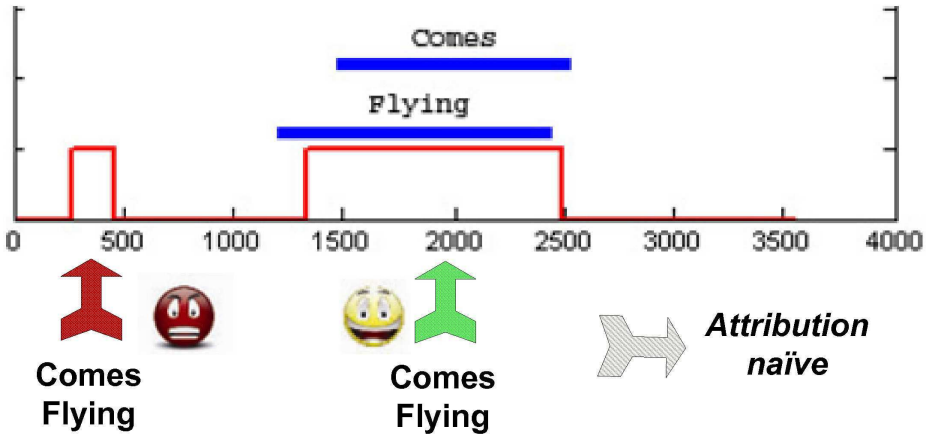

(c) Tamer of Wild Horses



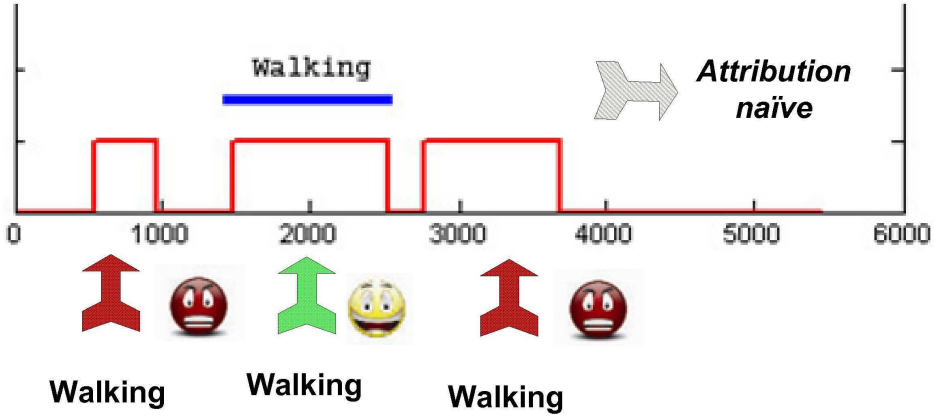

(d) The flying man

FIGURE 5.20 – Activité locale mesurée (rouge) et action (terme(s)) mise en scène dans la séquence vidéo (bleu) avec en abscisse le numéro de l'image dans la séquence.

à la mesure d'activité locale quasiment toutes les scènes d'activité décrites par le synopsis. Ce résultat globalement bon laisse cependant un certain nombre de situations où la présence d'une action décrite dans le synopsis n'a pas été confirmée par l'analyse d'image (voir la figure 5.21.c). Ceci se produit quand l'action relevée dans le synopsis se traduit par une activité peu marquée dans la séquence (par exemple un personnage qui court lentement). Par contre la mesure de précision est très basse du fait que cette annotation des sous-séquences est naïve. En effet, l'analyse des images nous fournit des sous séquences où l'activité est élevée et où l'action mise en scène n'est pas celle attribuée naïvement par le système (voir par exemple la première sous-séquence d'activité sur la figure 5.21.a qui est représentée par un smiley insatisfait). Cela se produit également lorsque l'activité mesurée dans les séquences d'images n'est pas confirmée par la présence d'un verbe d'action dans le synopsis, ce qui a




(a) Le moine et le poisson




(b) L'homme aux bras ballants

FIGURE 5.21 – Classification automatique des segments d'action

pour conséquence d'augmenter le nombre de « Faux Positifs » et donc de diminuer la mesure de précision.

## 5.5 Conclusion

Dans ce chapitre consacré à la caractérisation des films d'animation à partir de l'utilisation conjointe des informations issues de l'analyse des images et de l'analyse des textes, nous avons présenté deux approches applicatives pour la fusion d'information entre le texte et l'image. Dans cette étude nous avons d'abord présenté une caractérisation globale des films d'animation au travers des atmosphères dégagées par le film et le texte. Cette caractérisation de l'atmosphère dramatique est basée sur une fusion par expertise implémentée grâce aux systèmes flous. La fusion d'information entre ces deux sources a permis de montrer l'intérêt d'une telle approche pour diminuer l'incertitude de l'information fusionnée. En effet, une classification des genres dramatiques à partir de cette information fusionnée a permis d'obtenir de très bons résultats de précision. Enfin la comparaison de notre système de fusion avec d'autres approches nous a permis de valider ses performances mais surtout de montrer son intérêt qui est d'être complètement explicable. Ce dernier point est important dans notre domaine applicatif pour légitimer notre approche auprès des professionnels de l'animation.

Enfin, nous avons présenté une caractérisation locale des films d'animation à travers l'activité présente dans ces séquences. Dans cette étude nous avons utilisé deux types de descripteurs qui sont issus des péri-textes et des images. Notre approche utilise ces descripteurs conjointement en essayant de rapprocher une description de bas niveau sémantique, qui est l'activité dans une séquence d'animation, d'une description de haut niveau sémantique, qui est une description textuelle de cette activité. Cette étude **qualitative** nous permet de montrer que les termes dénotant un mouvement correspondent très souvent à une ou des sous-séquence(s) vidéo où l'activité locale est importante. En outre ces études qualitatives nous permettent d'envisager l'ajout de descripteurs et d'informations complémentaires mais nous permettent aussi d'envisager un passage à l'échelle et d'expérimenter notre approche sur l'ensemble de la base des films d'animations (dès obtention des droits d'exploitation) afin d'en extraire des règles et des résultats réellement significatifs. De plus, notons que la méthode de caractérisation de l'activité présentée ici est fortement liée au domaine de l'animation et que sa généralisation à des vidéos plus "conventionelles" impliquerait d'utiliser d'autres descripteurs image comme par exemple la détection de visage ou d'objets particuliers.

# Quatrième partie

# Conclusion

Chapitre 6

# Conclusions et Perspectives

## 6.1 Conclusions

Dans ce mémoire, nous nous sommes intéressés à la caractérisation automatique de séquences vidéo, et plus particulièrement à la caractérisation des films du festival d'animation d'Annecy. Pour être performante, l'indexation de telles données nécessite d'atteindre un niveau de description sémantique. L'apport original de notre travail se situe dans la nature des informations extraites et utilisées pour caractériser les documents vidéo. En effet, si l'utilisation de descripteurs issus de l'analyse des images ou du son est devenue incontournable pour la caractérisation de séquences vidéo, l'utilisation du texte et en particulier de péri-textes est beaucoup plus limitée. Les travaux qui proposent d'utiliser des informations textuelles connexes aux films sont rares. Aussi, ce travail propose d'utiliser conjointement les deux sources d'information que sont les séquences d'images et les synopsis des films. L'utilisation de ces textes et des informations descriptives de niveau sémantique élevé qu'ils contiennent permet de compléter et d'élever le niveau sémantique des informations issues des images. Cette approche multimodale situent nos travaux à la croisée d'un certain nombre de disciplines scientifiques : traitement de l'image, traitement automatique des langues, ingénierie des connaissances, fusion d'information. Même si notre démarche a été construite en tentant de lui donner un caractère générique, l'application de cette approche s'est appuyée sur les spécificités du domaine envisagé, celui du film d'animation. Le contexte local a joué un rôle important dans le choix de ce domaine. En effet, Annecy, avec son Festival International du Film d'Animation, est devenu depuis plus de quarante ans une référence mondiale dans le monde de l'animation. On peut également notre que l'industrie de l'animation a connu ces dernières années un essor important, en particulier grâce à l'évolution des techniques de synthèse d'images 3D. Dans ce contexte, nos travaux constituent une des premières démarches s'intéressant à la caractérisation sémantique des films d'animation par fusion multimodale. Cette problématique a été abordée en utilisant une analyse à deux niveaux :

- Une analyse bas niveau où des informations de couleurs et d'activité sont extraites des séquences d'images. Ces informations d'un niveau sémantique assez bas constituent une description globale des caractéristiques de la séquence d'animation.

- Une analyse haut niveau où des informations descriptives sont extraites des textes et en particulier des synopsis. Ces informations d'un niveau sémantique élevé constituent une description précise de la séquence d'animation proche des concepts manipulés par

l'Homme.

*L'analyse des séquences d'images* abordée ici est un prolongement des travaux réalisés dans la thèse de Bogdan Ioenscu [Ionescu, 2007] sur la caractérisation symbolique des films d'animation. Des caractéristiques liées à la couleur comme la diversité, l'utilisation de couleurs foncées, etc. sont extraites d'une analyse statistique des images. Cette approche est basée sur l'utilisation de dictionnaires couleurs, qui permettent une caractérisation sémantique globale de la signature couleur de la séquence vidéo. Une amélioration de l'analyse des séquences d'images a été proposée dans nos travaux pour lui donner un caractère générique nécessaire pour traiter l'ensemble des films. En effet, la méthode initiale est basée sur la détection du découpage temporel du film en séquences. Or ce découpage habituellement utilisé dans les longs métrages ou les films naturels n'est pas toujours mis en œuvre dans les très courtes séquences vidéo. De plus, notre approche basée sur un algorithme à accumulation d'erreur permet de mesurer l'activité et le rythme de la séquence vidéo de façon plus complète que la méthode proposée initialement dans [Ionescu, 2007]. En effet, cette mesure est liée aux changements de contenu s'opérant dans les images et permet ainsi de mesurer l'activité intra et inter plans.

*L'analyse des textes* est abordée ici suivant deux approches qui donnent naissance à deux caractérisations différentes. Une première approche statistique utilise un ensemble de dictionnaires thématiques qui permettent de mesurer l'intensité du thème à être présent dans le texte. Cette mesure permet de repérer des atmosphères dégagées par le récit comme l'atmosphère dramatique. Cette information permet de caractériser le film de façon globale et permet de déterminer le genre du film. Une deuxième approche basée sur les méthodes d'extraction d'information permet à partir des analyses lexicale, syntaxique et sémantique d'extraire un scénario actanciel. Ce modèle de représentation de l'information permet de modéliser l'action mise en scène dans le film.

Finalement, *la caractérisation des films* est envisagée à partir de l'analyse conjointe de ces deux sources d'information et suivant deux niveaux d'abstraction. La caractérisation **globale** du film est abordée au travers de l'analyse du genre "drame". En effet, la fusion du texte et de l'image permet de tirer profil de la complémentarité de ces deux sources d'information pour caractériser des atmosphères dramatiques. L'information d'appartenance au drame est obtenue à partir des descripteurs images et de leur fusion. Cette fusion entre les informations colorimétrique et d'activité permet d'obtenir de nouveaux concepts liés à cette thématique. Cette information issue de l'image est fusionnée à l'intensité thématique mesurée à partir du texte afin de diminuer l'incertitude des informations et permettre ainsi de classifier les films comme dramatique avec une bonne précision. Cette approche, mise en place à travers la réalisation d'un système flou, permet d'obtenir de bons résultats et aboutit à une caractérisation du drame avec une bonne certitude.
Le deuxième niveau de caractérisation des films est effectué au niveau local. Cette caractérisation a pour but de repérer dans le temps (sur le support vidéo) les éléments du scénario actanciel, afin de retrouver et nommer localement ses éléments (personnage(s), action, contexte) dans les sous-séquences vidéo. Le travail réalisé est une première étape de cette caractérisation dont l'objectif est la vérification de l'alignement des deux sources d'informations. Cet alignement est réalisé grâce à l'utilisation conjointe des éléments textuels du scénario actanciel portés par les verbes d'action et la mesure d'activité locale issue de l'image. Nous avons montré sur la base de tests que cet alignement est réalisable et permet de retrouver

généralement les sous-séquences décrites par le scénario actanciel.

## 6.2 Perspectives

Dans ces travaux nous avons montré la complémentarité des sources d'information que sont les images et le texte. Cette complémentarité permet une caractérisation multimodale des films. Cependant plusieurs améliorations peuvent être envisagées :

- Il est envisagé un passage à l'échelle. En effet, les expérimentations (hors texte) présentées dans ce manuscrit sont réalisées sur un sous ensemble de la base vidéo (107 films), contrainte imposée par un problème de droits d'exploitation. Cette base de test reste trop petite pour obtenir des résultats statistiques réellement significatifs. Une solution envisageable pour contourner ce problème de droits est d'effectuer directement les calculs et les analyses *in situ* (dans les locaux de CITIA). Cela nécessite, au niveau technique, le développement d'une plateforme robuste et paramétrable à distance qui puisse analyser les films et envoyer les résultats. Cela nécessite également la mise en place d'une convention CITIA / Université de Savoie, convention pour laquelle de premières discussions sont déjà engagées.

- Il est envisagé d'ajouter de nouveaux descripteurs caractérisant le genre. C'est par exemple la prise en compte de la modalité son. En effet, lorsque cette dernière est présente dans la séquence, elle pourrait jouer un rôle important dans la caractérisation sémantique du contenu des films d'animation (présence de dialogues, intensité et rythme de la musique, silences, bruits, etc.). Ces informations peuvent servir pour la caractérisation des atmosphères [Trohidis *et al.*, 2008] et notamment pour retrouver d'autres films dramatiques où les informations de couleurs et le synopsis n'apportent pas conjointement les informations caractéristiques du drame (amélioration de la mesure de rappel). Elles peuvent servir également pour la caractérisation locale de la séquence d'animation en permettant l'alignement des modalités. L'information de présence de dialogues permettrait de retrouver les sous-séquences où des personnages sont présents.

- Il est envisagé d'ajouter de nouveaux descripteurs images pour retrouver localement les mouvements des personnages. En effet, des travaux sont en cours sur l'extraction et la caractérisation des points d'intérêt spatio-temporels. Ces points permettent de repérer dans les séquences d'images les points ou régions qui se déplacent dans les images. Ces points correspondent généralement à des personnages en action ou à des objets animés d'un mouvement. L'intérêt d'un tel descripteur est double. En effet, la mesure globale de l'activité de ces points est une mesure de l'action probablement plus précise que celle présentée dans ces travaux (basée sur les images clefs) car directement liée à l'action des personnages. Cette mesure de l'activité permettrait probablement un meilleur alignement des sources. De plus, le deuxième intérêt serait une caractérisation par l'image (direction, sens, rapidité, etc.) des actions décrites par le scénario actanciel. Le passage par une étape d'apprentissage permettant de lier ces caractéristiques à des concepts haut-niveau permettrait de créer un réseau sémantique où les descriptions bas niveau (rapidité, direction, etc.) seraient associés aux termes. Par exemple, la différence dans la direction du mouvement permettrait la distinction entre **sauter** et **marcher**, et la différence dans l'amplitude du mouvement distinguerait **courir** et **marcher**. Un tel

réseau permettrait par la suite de faire automatiquement le lien entre la description du scénario actanciel et la mesure de l'action dans l'image. Imaginons que dans une sous-séquence on détecte deux objets en mouvement. Le premier mouvement est caractérisé par une direction verticale alors que le deuxième est caractérisé par une direction horizontale. De plus, si le synopsis fait référence aux actions "sauter" et "courir" comme dans : « Marie saute sur le trampoline pendant que Jean court après le ballon » alors ce réseau sémantique permettrait de nommer automatiquement dans l'image le premier mouvement comme étant l'action "sauter" et de nommer automatiquement dans l'image le deuxième mouvement comme étant l'action "courir". Dans cet exemple on pourrait également nommer automatiquement les personnages.

– Enfin, il est envisageable d'ajouter de nouveaux descripteurs images pour retrouver localement les personnages, qui sont des éléments fondamentaux pour la caractérisation du contenu des films. Une des difficultés majeure dans la reconnaissance des personnages d'animation vient de l'extrême variabilité des caractéristiques couleurs, textures, et formes (voir le chapitre consacré à l'analyse des images). En effet, la magie de l'animation est de pouvoir donner la vie à des objets qui par nature sont inanimés. Cependant, ces "personnages" animés ont presque systématiquement une caractéristique descriptive commune : ils ont des yeux. Ces sont d'ailleurs ces yeux qui leur confèrent la qualité de personnage. Ainsi la mise en place d'un détecteur capable de retrouver des yeux dans une région de l'image animée d'un mouvement permettrait très certainement de retrouver les personnages du film.

# Cinquième partie

# Annexes

Annexe A

# Les techniques d'animation

C'est l'utilisation de techniques particulières qui distingue fondamentalement le cinéma d'animation du cinéma classique. Ces techniques peuvent être regroupées suivant certaines caractéristiques dont voici une liste non exhaustive.

## A.1 Le dessin animé :

Cette technique d'animation est probablement la plus connue parmi les techniques utilisées dans l'animation. Traditionnellement la version dite "plastique" est constituée de décors peints sur papier, sur carton ou sur toile (dont la colorisation varie : gouache, acrylique, aquarelle ou autre) et des personnages et objets mouvants dessinés sur feuilles puis encrés et gouachés sur celluloïdes. Le cell, ou cellulo, ou celluloïd est une feuille plastique transparente d'acétate de cellulose sur laquelle sont peints à la main les différents éléments d'un dessin animé. Les cellulos sont ensuite superposés et, grâce à leur transparence, il est ainsi possible de créer des scènes complexes sans tout redessiner à chaque fois (toute partie immobile n'ayant pas à être redessinée). Cela permet également de créer des effets de perspective de mouvement. Par exemple lors d'un travelling les techniciens font défiler les calques à des vitesses différentes, d'autant plus lentes que le plan est éloigné. Il existe aussi d'autres techniques utilisant d'autres matériaux et techniques de colorisation (animation de personnages sur des feuilles de papiers, coloriés à la craie ou aux crayons de couleur par exemple). La production des grands studios américains (Disney, Warner, MGM, Hanna-Barbera) a été essentiellement réalisée à partir de cette technique (voir figure A.1). Les œuvres de pionniers comme le Français Émile Cohl (Fantasmagorie, 1908) et l'Américain Winsor McCay (le créateur de Gertie le dinosaure, 1914) ont été réalisées avec cette technique. Pour de plus amples informations voir [Williams *et al.*, 2003].

## A.2 Animation d'objets 2D :

Dans cette technique d'animation, l'auteur va utiliser des objets pour composer un plan (2D). La prise de vue est réalisée image par image (souvent en vue de dessus) et c'est le déplacement et la composition de ces objets qui va donner une illusion de mouvement. Les matériaux les plus fréquemment utilisés sont :

**-Le Sable :** Afin de réaliser une image en mouvement, cette technique utilise les propriétés plastiques du sable. Le principe général (dont la prise de vue est réalisée par le dessus)

FIGURE A.1 – Les dessins animés : conception et dessin

est l'utilisation d'une table lumineuse sur laquelle est disposée une couche de sable. Par déplacement de celui-ci, les couches de sable deviennent plus ou moins épaisses et par conséquent l'effet d'ombre et de lumière plus ou moins marqué (voir figure A.2 (a)).

**-Le Papier découpé :** Cette technique est une des premières techniques utilisées dans les films d'animation mais aussi l'une des plus économiques. Le mouvement peut être obtenu en remplaçant divers éléments découpés ou en animant des personnages composés d'éléments articulés. Cette technique est récemment réapparue dans la désormais célèbre série ***"South Park"*** ainsi que dans une publicité pour un opérateur téléphonique mobile français (voir figure A.2 (b)). L'animation de photographies découpées constitue une variante de l'animation de papier découpé traditionnelle.

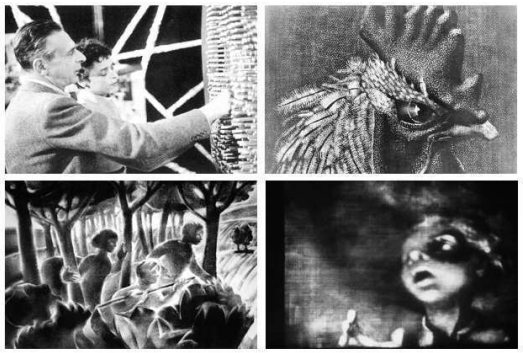
**-L'écran d'épingles :** Cette technique inventée par le Français **Alexandre Alexeïeff** consiste à utiliser un écran blanc placé verticalement et percé de centaines de milliers de trous ($\simeq$ 240 000 trous) chacun traversé par une épingle noire rétractable dépassant de la surface de l'écran. Une lumière est projetée de biais des deux cotés de l'écran ce qui fait que les ombres des épingles rendent l'écran noir. En faisant varier le degré d'enfoncement de certaines épingles (de façon à former un dessin en relief), leurs ombres raccourcissent et ainsi le gris remplace le noir. Les épingles qui sont totalement enfoncées ne laissent plus d'ombre et font donc apparaître le blanc de l'écran. L'esthétique de l'image ainsi formée est fidèle au graphisme et aux dégradés de gris présents dans les gravures. C'est en 1933 qu'**Alexeïeff**, aidé de sa collaboratrice **Claire Parker**, termine ***"Une nuit sur le mont Chauvei"*** avec cette technique (voir figure A.2 (c)).

## A.3 Animation en volume (objets 3D) :

Une scène (en général constituée d'objets) est photographiée image par image. Entre chaque image, les objets de la scène sont légèrement déplacés. Lorsque le film est projeté à une vitesse normale, la scène semble animée. L'animation en volume a en commun avec le cinéma classique d'exiger du cinéaste qu'il tienne compte d'éléments de mise en scène comme lors des prises de vues réelles. En effet, le choix de l'objectif, les mouvements de caméra, la profondeur de champ, l'éclairage, et les rapports spatiaux entre les éléments ne sont pas virtuels, comme en dessin animé, mais plutôt réels, comme dans les films de fiction avec acteurs. Parmi cette technique on retrouve :

(a) : Animation de Sable


(b) : Éléments découpés



(c) : Ecran d'épingles

FIGURE A.2 – Animation d'objets 2D

**-Les Marionnettes :** L'animation de marionnettes ou animation de poupées est fortement inspirée du théâtre de marionnettes. Cette tradition d'Europe centrale explique l'implantation initiale de cette technique, dans cette partie du monde (***"La vengeance de l'opérateur cinématographique" (1912)*** du Russe **Ladislas Starewitch**) (voir figure A.3 (a)).

**-La Pâte à modeler ou "claymation" :** Cette fois-ci les objets sont sculptés dans de la pâte à modeler. Parmi les réalisations les plus connues de ces dernières années figurent ***Wallace et Gromit*** et ***Chicken Run*** (voir figure A.3 (b)).

**-La Pixilation :** Dans cette technique ce sont des acteurs réels ou des objets qui sont filmés image par image. Ce terme utilisé la première fois par **Norman McLaren** signifie en anglais être dirigé/ensorcelé par un pixy, sorte de fée ou de lutin (rien à voir avec les pixels). Il utilisa cette technique dans ***"Voisins" (1952)*** (voir figure A.3 (c)).

## A.4 Animation numérique :

Cette technique est la plus récente des techniques d'animation. Elle consiste en la création assistée par ordinateur, ce qui implique que tout ou partie de l'animation est réalisée numériquement. En animation 2D numérique le crayonné est souvent fait sur papier ou calque. Le coloriage et la composition [1] sont ensuite généralement faits sur ordinateur après numé-

1. La composition ou "composting" consiste à assembler toutes les couches des décors, des personnages et à réaliser les effets de caméra, etc. pour en faire un plan unique

(a) : Marionnettes

(b) : Pâte à modeler

(c) : Pixilation

FIGURE A.3 – Animation en volume

risation du crayonné. Cependant, il est assez fréquent que les décors ou que les objets et personnages animés soient déjà coloriés, ou que tout soit complètement numérique (images de synthèse via une tablette graphique par exemple). Finalement, Les possibilités de mixage entre animation numérique, effets spéciaux numériques et animation traditionnelle sont quasiment infinies. En animation 3D numérique les possibilités de mixage sont les même qu'en 2D et l'approche cinématographique est équivalente à l'animation en volume mais cette fois ci dans un monde virtuel.

Depuis quelques années, l'utilisation de l'ordinateur est devenu un outil complémentaire dans les mains des artistes utilisant les techniques traditionnelles. Ainsi, depuis les années 1990, l'apport de l'informatique se généralise et se diversifie dans le cinéma d'animation faisant qu'aujourd'hui, cet outil est devenu incontournable dans la réalisation des films. L'objectif de cette section n'est pas de faire la liste exhaustive des techniques d'animation qui est quasi infinie, mais plutôt de marquer la spécificité des films présentés au FIFA par la description de quelques techniques d'animation assez répandues. Parmi l'éventail de possibilités que les auteurs ont à disposition pour faire passer leurs intentions artistiques, l'usage de la couleur dans les séquences d'animation est lui aussi assez spécifique.

Annexe B

# La base d'animation de CITIA

Nous presentons ici les caractéristiques de la base de films d'animation. L'ensemble des fiches d'inscription des films inscrits au FIFA constitue une base de données textuelles. Cette base de données issue de Animaquid contient 5804 entrées dont voici quelques caractéristiques statistiques.

## B.1 Répartition des films en fonction de l'année d'inscription

Tout d'abord, le FIFA est un événement annuel qui est né au début des années 60. Lorsque l'on regarde la répartition des films disponibles suivant l'année d'inscription au festival (figure B.1) on voit que depuis les années 80, il y a une augmentation du nombre de films inscrits et sélectionnés. Ceci montre que le FIFA devient un événement incontournable dans le domaine de l'animation mais également que ce domaine cinématographique connaît un succès dans des domaines variés depuis un peu plus 20 ans.

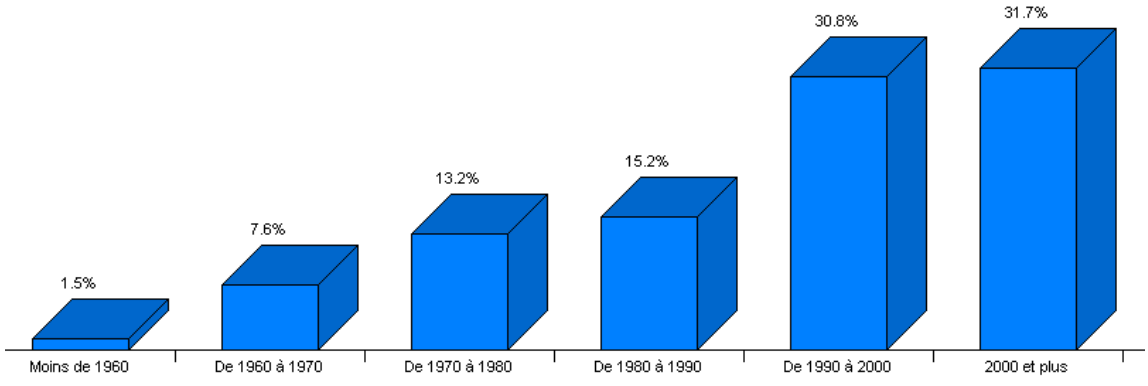



Figure B.1 – Répartition des films suivant l'année de production

## B.2 Répartition en fonction de la durée des films

Une des caractéristiques importante de ces films concerne la durée de la séquence vidéo. La durée moyenne des films de la base est de 7.75 minutes (avec un écart-type de 11.29

minutes). On voit sur la figure E.13 que la majorité des films (80%) ont une durée inférieure à 10 minutes. Les films de la base sont donc des films courts et cette caractéristique constitue une information *a priori* très importante, notamment pour l'analyse des images et l'analyse des synopsis.

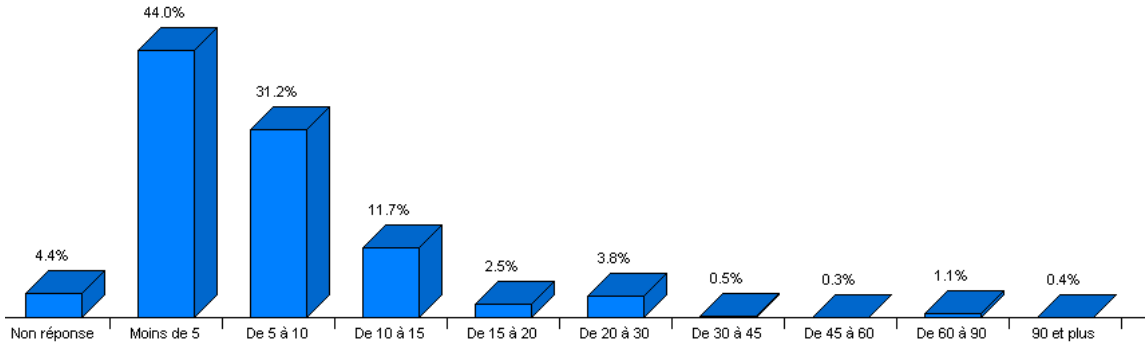



FIGURE B.2 – Répartition des films suivant la durée exprimée en minute

## B.3 Répartition des films par pays de production

Sur la figure B.3 on voit la répartition des films par pays de production (seuls les plus représentés y sont figurés). Ce que l'on voit clairement, c'est que les pays occidentaux sont très présents dans cette manifestation. On voit également apparaître la dimension internationale avec les pays comme les États-unis, le Canada, ou le Japon, etc. Mais ce qui est frappant c'est que 50% des films inscrits au festival sont soit Français, Anglais, ou Américains.

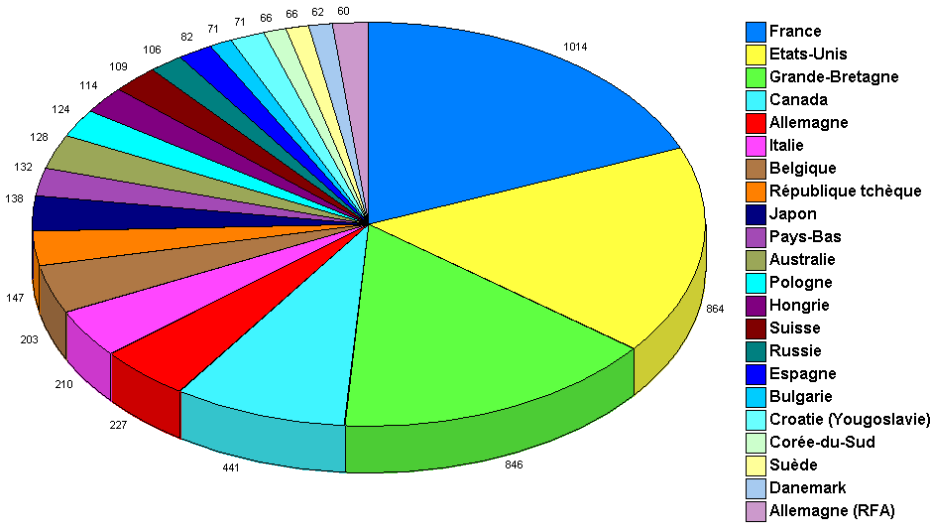



FIGURE B.3 – Répartition des films suivant le pays de production *(les valeurs numériques sont les effectifs)*

Quand on regarde la contribution des pays au cours du temps on obtient la figure B.4

obtenue par AFC dont l'axe X (abscisse) représente 57% de la variance expliquée et dont l'axe Y (ordonnée) représente 17% de la variance expliquée. On voit sur cette figure qu'il y a eu un apport significatif de films en provenance de Grande Bretagne durant la période de 1990 à 2000 et en provenance des États-Unis durant la période de 1970 à 1980. La part française se situant plus dans la période actuelle (années 2000). En outre il est intéressant de constater que, dans les débuts du festival, une part significative des films venait de l'Europe de l'Est.

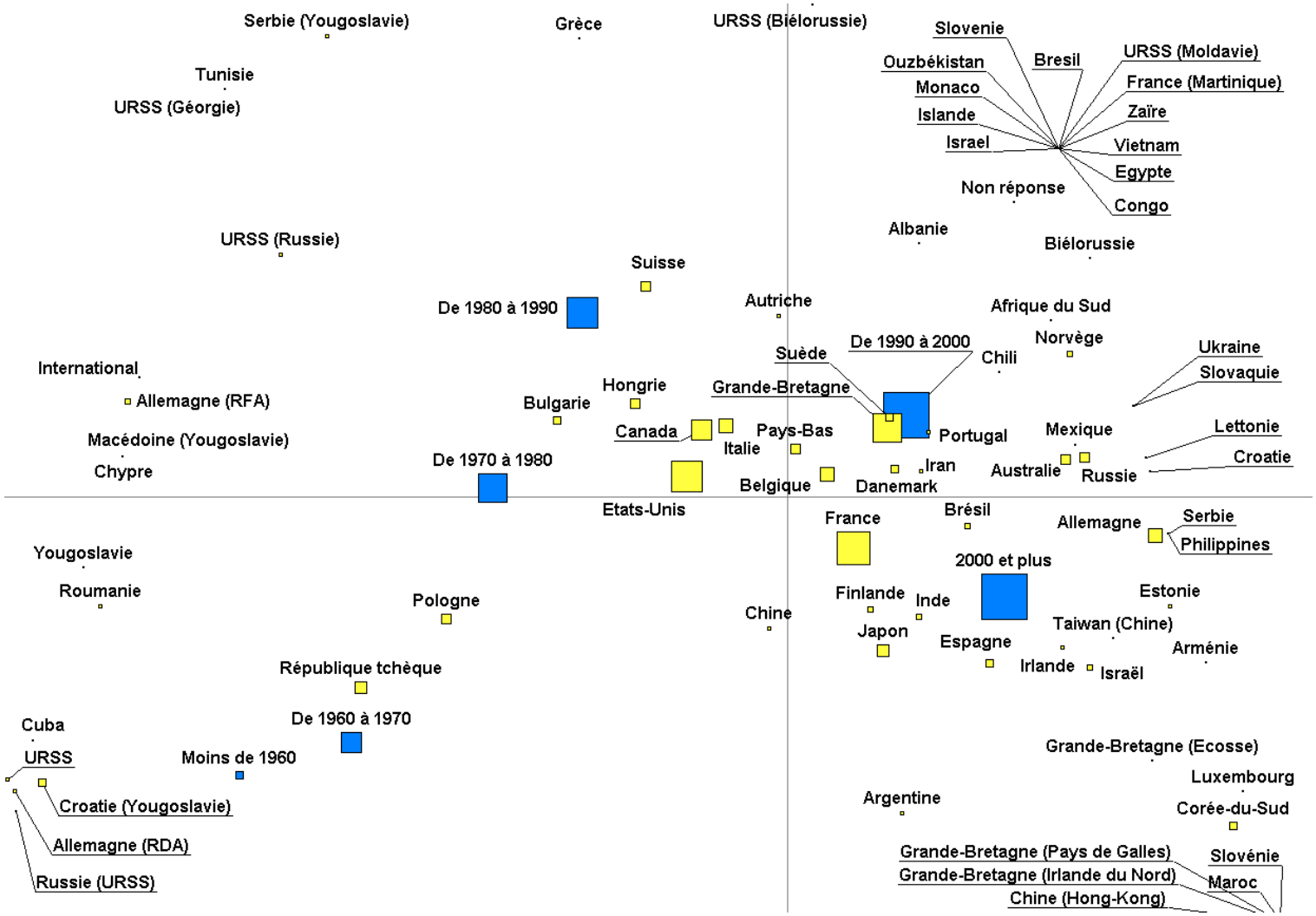



FIGURE B.4 – Carte AFC : répartition des films par pays de production et année d'inscription au festival *(la surface des carrés est proportionnelle aux effectifs, les carrés bleus sont les périodes de production, les carrés jaunes sont les pays de production)*

## B.4 Répartition des films suivant le public visé

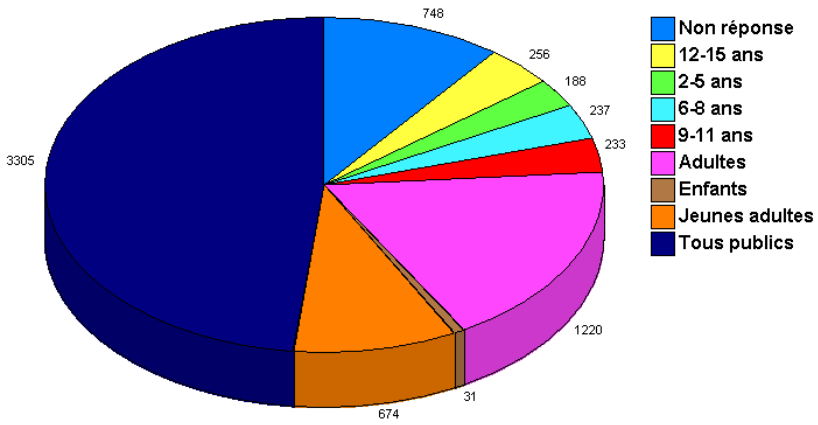



FIGURE B.5 – Répartition des films suivant le public visé *(les valeurs numériques sont les effectifs)*

Sur la figure B.5 on voit la répartition des films suivant le public visé. Il apparaît clairement que le contenu des œuvres est une des caractéristiques des films inscrits à Annecy. En effet 25% de ces films s'adressent à un public d'adultes. On est donc dans un type de contenu assez éloigné des "classiques" de chez Disney car seulement 15% des films s'adressent à un public d'enfants.

## B.5 Répartition des films suivant la technique d'animation

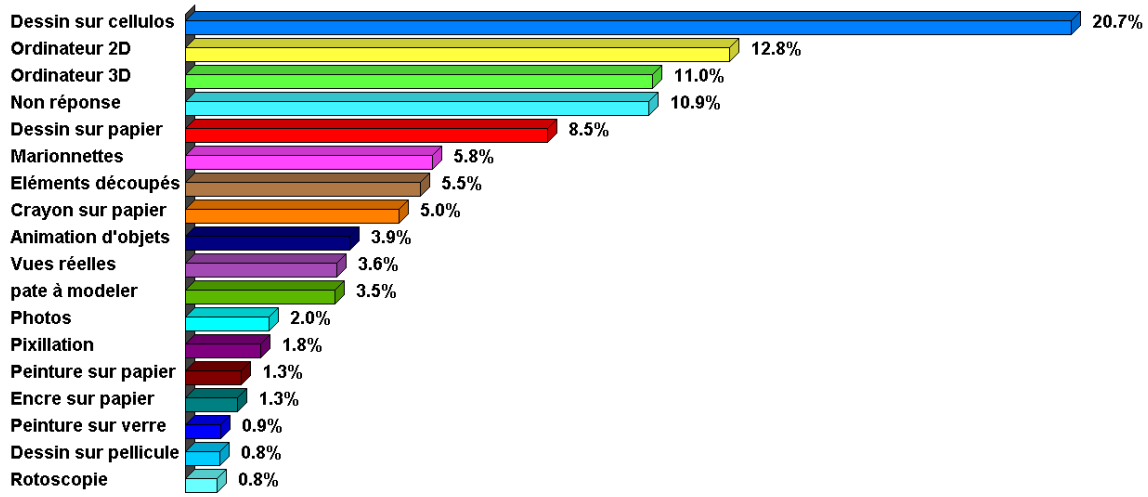



FIGURE B.6 – Répartition des films suivant les principales techniques d'animation *(Les pourcentages sont calculés sur les 5804 entrées de la base)*

Parmi les 70 techniques d'animation présentes dans la base textuelle, seules les techniques

significatives ont été conservées sur la figure B.6. Ainsi n'apparaissent que les techniques d'animation dont l'occurrence est supérieure à 60. Le dessin sur celluloïds, une des techniques les plus connues, a été significativement utilisé durant les années 80 (voir figure B.7) et demeure la technique majoritairement utilisée dans la base de CITIA (voir figure B.6). L'animation numérique, qui est une technique "jeune" (voir figure B.7) connaît un engouement ces dernières années et est également très présente dans la base.

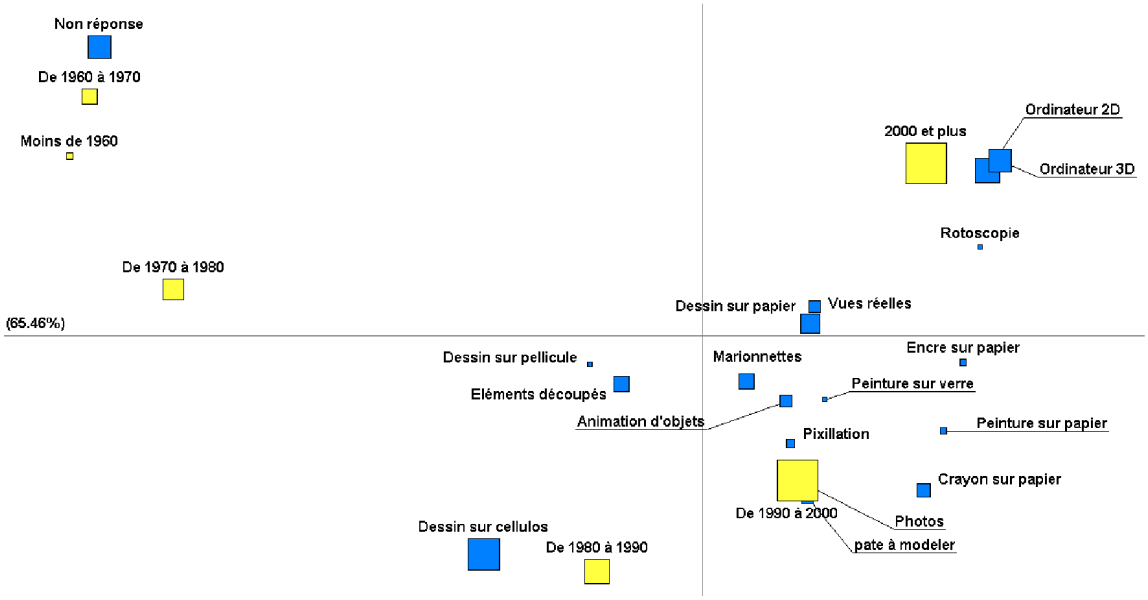



FIGURE B.7 – Carte AFC : répartition des films par technique d'animation et année d'inscription au festival *(la surface des carrés est proportionnelle aux effectifs, les carrés jaunes sont les périodes de production, les carrés bleus sont les techniques d'animation)*

## B.6 Répartition des films suivant le genre d'animation déclaré

Malheureusement, les genres d'animation (qui sont le reflet du contenu du film) souffrent d'une très grande hétérogénéité. Dans la base de données, ce champ est constitué de beaucoup de catégories qui ne sont pas, ou ne peuvent pas être associées à un genre d'animation comme défini dans l'ontologie des genres. Bien que le quart de la base n'a pas un genre d'animation bien défini, on remarque que les films humoristiques sont les plus présents au festival. Cependant le thème du drame représente une part importante des films dont le genre est attribué ($\simeq 11\%$).

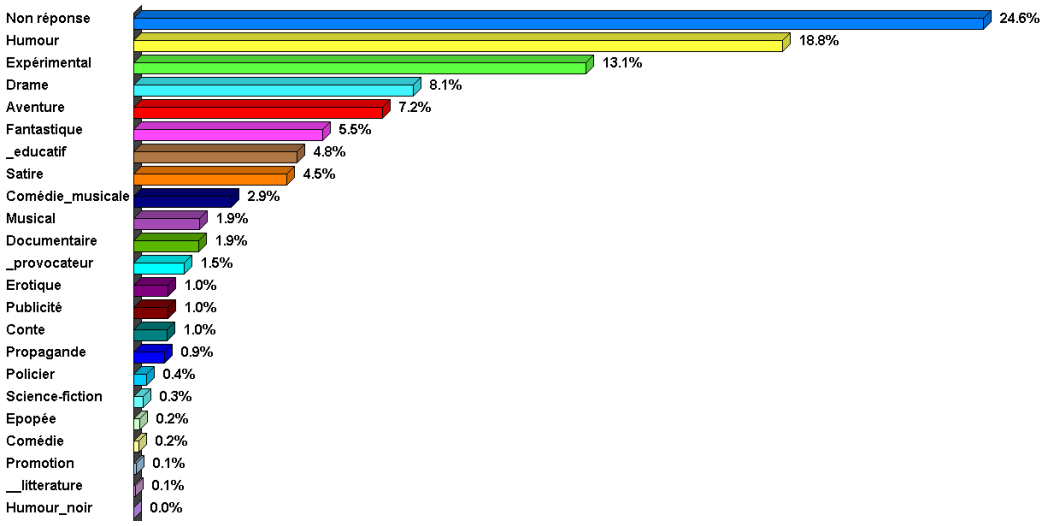



FIGURE B.8 – Répartition des films suivant les genres *(Les pourcentages sont calculés sur les 5804 entrées de la base. NB : Le genre humour_ noir n'apparaît que 2 fois)*

## B.7 Répartition des synopsis suivant le nombre de mots

La caractéristique des synopsis dont nous disposons dans cette base textuelle est leur longueur. En effet les figure B.9 et B.10 montrent la répartition des synopsis en fonction du nombre de mots qu'ils comportent. Pour les synopsis français le nombre de mots moyen est 24 mots avec un écart-type de 15. Pour les synopsis anglais le nombre de mots moyen est 22 mots avec un écart-type de 14. Ainsi que ce soit en anglais ou en français les textes dont nous disposons sont des textes courts (68% des synopsis ont entre 10 et 40 mots) avec une grande variabilité (en nombre de mots).

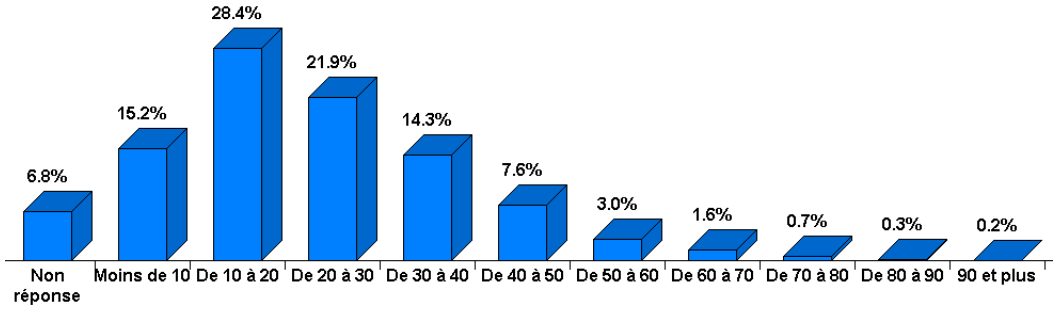



FIGURE B.9 – Répartition des synopsis français suivant le nombre de mots qu'ils contiennent

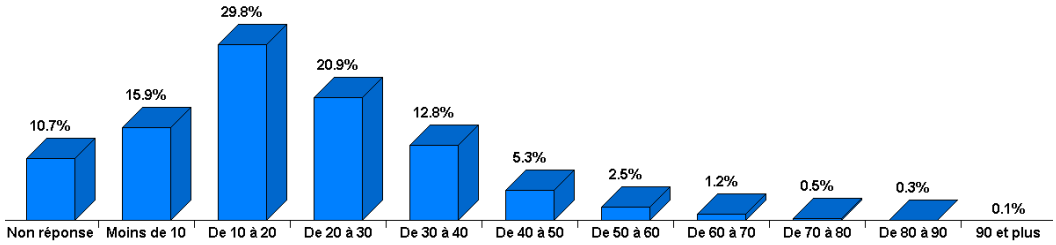



FIGURE B.10 – Répartition des synopsis anglais suivant le nombre de mots qu'ils contiennent

Annexe C

# Tests et résultats de l'analyse d'image

Pour valider notre approche, nous avons testé notre algorithme sur une base de 10 films d'animation dont l'auteur [Bouillot, 2008] a créé pour chaque film un résumé statique (voir figure C.1 (a)). Ces images apparaissent pour l'auteur comme des images essentielles pour caractériser son œuvre. Ces résumés constituent donc notre vérité terrain.

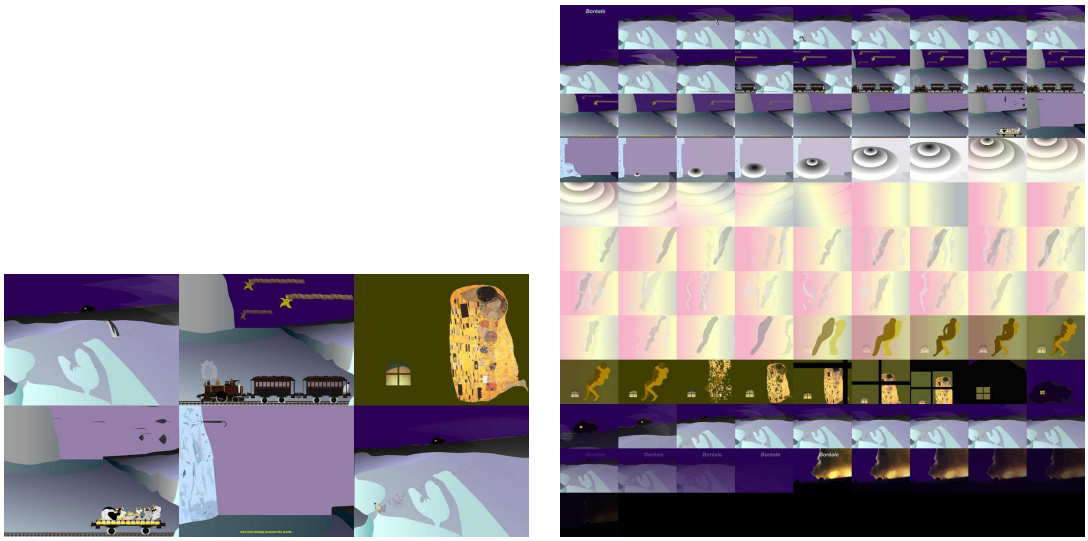
Pour exploiter cette dernière nous avons repéré dans chaque film et pour chaque image fournie le début et la fin de la sous séquence dans laquelle apparaît cette image, obtenant ainsi, ce que nous appellerons les plages de la vérité terrain. Notre objectif est d'obtenir un condensat d'image le plus proche possible de la vérité terrain.

## C.1 Le choix de la méthode de comparaison des blocs

Pour valider le choix de la méthode de comparaison des blocs nous avons testé notre algorithme d'extraction d'image clefs sur la base des 10 films d'animation de Daniel Bouillot. Pour mesurer ses performances nous partons de l'hypothèse que les images clefs retournées par l'algorithme doivent être le moins redondantes tout en couvrant le maximum de passages et moments du film et en particulier les passages importants de la vérité terrain. Nous avons donc vérifié si les numéros des images clefs retournées par l'Algorithme à Accumulation (AaA) appartenaient aux plages de la vérité terrain. De cette manière nous cherchons à vérifier si l'AaA permet d'obtenir un ensemble d'images clef au contenu le moins redondant tout en couvrant le maximum de passages et moments importants du film.

Pour évaluer les performances de la méthode proposée nous l'avons comparée à d'autres méthodes de comparaison entre blocs.

1. La première **(RGB)** est la méthode originale [Lu et Suganthan, 2004] basée sur le calcul d'une distance Euclidienne dans l'espace RGB. Le seuil étant fixé de manière empirique.
2. La deuxième est l'utilisation de distances de similarité **DeltaE(CIE 1976)** et **DeltaE(CMC)** définis dans l'espace colorimétrique CIE1976Lab par le Color Measurement Committee (CMC). Afin d'être en mesure de comparer les blocs, le seuil est fixé à 7 comme préconisé dans [Mojsilovic, 2005].
3. La troisième **(Name267)** est la méthode proposée dans ce travail en utilisant les noms des couleurs. Nous avons utilisé en réalité 2 comparaisons. Une comparaison « pleine »



(a) Vérité Terrain (b) Résultats

FIGURE C.1 – Exemple sur un film d'animation (*Boréale, (3197 images)*) : (a) Le résumé statique fourni par l'auteur, (b) Le résumé à la sortie de l'algorithme à accumulation.

avec les 267 couleurs du ISCC-NBS et une comparaison « réduite » aux 31 couleurs de base de ces 267 couleurs **(Name31)**.

4. Enfin, une expérience témoin qui n'est autre qu'un tirage aléatoire **(Rand)** des images clefs a été également utilisée pour relativiser les résultats.

Les résultats sont présentés à travers la précision $P$ et le rappel $R$, calculés sur chaque film et moyennés sur l'ensemble des 10 films étudiés, dont voici une synthèse :

| **Méthode** | Name267 | DeltaE76 | DeltaECMC | Name31 | RGB | Rand |
|---|---|---|---|---|---|---|
| **Précision** | 0.31 | 0.32 | 0.28 | 0.26 | 0.23 | 0.15 |
| **Rappel** | 0.93 | 0.90 | 0.81 | 0.73 | 0.57 | 0.37 |

### C.1.1 Discussions

Notre objectif est d'obtenir un ensemble d'image les moins redondantes possibles tout en couvrant le maximum de passages du film et notamment ceux qui sont considérés comme importants pour l'auteur. La mesure de précision va être affectée par le nombre d'images clefs non comprises dans les plages de la vérité terrain. Elles correspondent aux images comprises dans des passages *"moins importants"* du film qui sont forcement extraites par l'AaA. Cette mesure est surtout affectée par le nombre d'images redondantes extraites de ces passages. La mesure de rappel, quant à elle, va être affectée par le nombre d'images comprises dans les plages de la vérité terrain mais non extraites par l'AaA. Cette mesure est liée à l'objectif de couvrir le maximum de passages importants du film.

On peut voir sur la figure C.2 que les méthodes du « color naming » (Name267) et de la distance dans l'espace Lab (DeltaE76) donnent de bons résultats. On peut remarquer que le

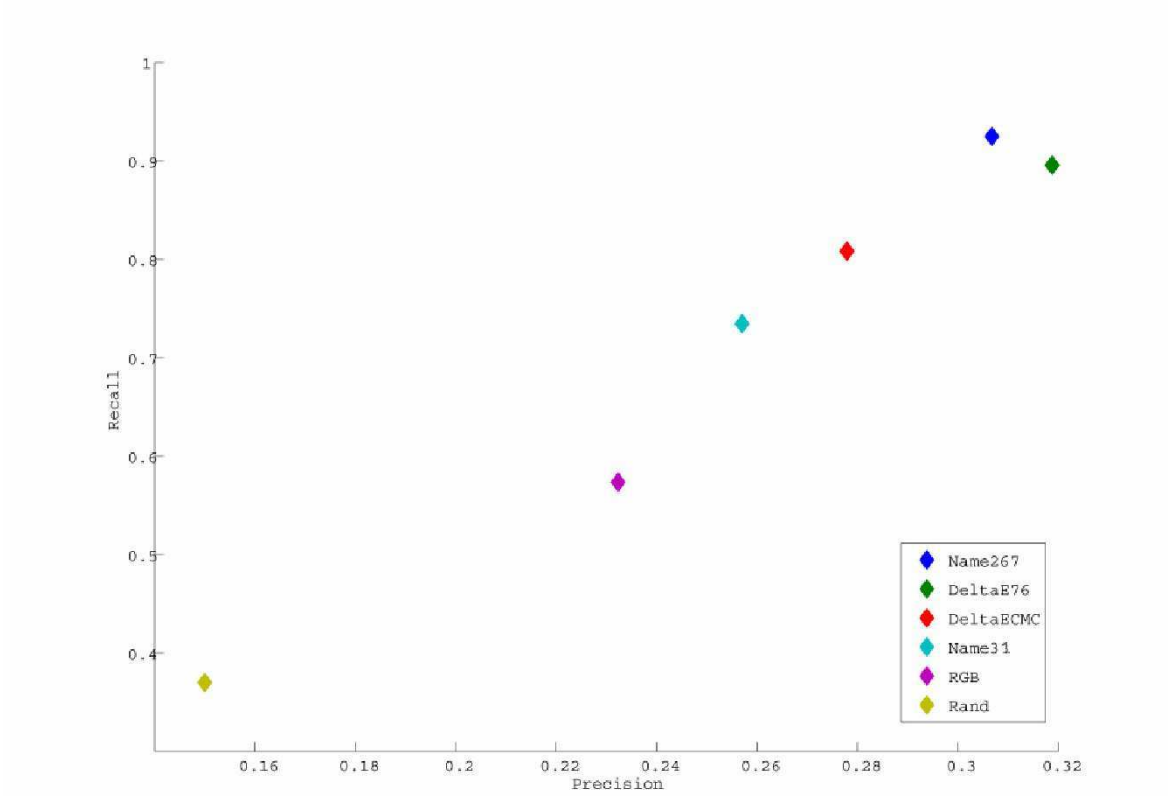



FIGURE C.2 – Graphique Précision Rappel en fonction des méthodes de comparaison des blocs

rappel est très bon alors que la précision est passable. Ceci vient du fait que le nombre de « *Key Frame* » par résumé de la vérité terrain est en moyenne de 4 images et que l'AaA réduit en moyenne la longueur du film à 3.5% de sa longueur originale. On a donc beaucoup d'image à la sortie de l'AaA (voir figure C.1) par rapport au nombre d'images de la vérité terrain. Les images de la vérité terrain ayant une forte probabilité d'être dans les images clefs retournées par l'algorithme on a donc un rappel élevé. De la même manière beaucoup de bruit se trouve dans le résultat ce qui explique pourquoi la précision est faible. Par contre la méthode qui consiste à calculer la distance dans l'espace RGB donne des résultats similaires à la méthode de sélection aléatoire. En réalité si cette dernière donne de bons résultats c'est parce que la couverture de la vérité terrain est assez importante. Les sous séquences correspondant à la vérité terrain (plages) représentent en moyenne 23% de la longueur du film.

## C.2 Le choix des distances dans la classification ascendante hiérarchique

La CAH demande la définition préalable d'une distance entre les individus à classifier (dans notre cas les individus sont des images). De nombreuses distances ont été utilisées afin

de juger de la similarité entre les images :

### C.2.1 Distance entre individus

1. La première distance est basée sur un **histogramme global** (équation C.1) dont le nombre de couleurs est réduit en utilisant une discrétisation uniforme de l'espace RGB. Ainsi le cube de l'espace RGB est divisé en 125 petits cubes et chaque couleur est remplacée par la valeur de la couleur du centre du cube auquel elle appartient [Ionescu *et al.*, 2005b].

$$d(m,n) = \sqrt{\sum_{i=1}^{Nc}(H_m(i) - H_n(i))^2} \tag{C.1}$$

   Pour chaque image $f_m$ et $f_n$ deux histogrammes couleur $H_m$ et $H_n$ normalisés sont calculés. Ensuite la distance Euclidienne (équation C.1) est calculée sur ces histogrammes, avec $i$ l'index de la couleur et $Nc = 125$ représentent le nombre de couleurs.

2. La deuxième est basée sur la **décomposition en bloc**.

$$d(f_i, f_j) = \frac{1}{N^2}\sum_{k=1}^{N^2}(DeltaE76(F_i(k), F_j(k)))$$

$$DeltaE76(P_x(L,a,b), P_y(L,a,b)) = \sqrt{(L_x - L_y)^2 + (a^*_x - a^*_y)^2 + (b^*_x - b^*_y)^2} \tag{C.2}$$

   Où $f_i$ et $f_j$ désignent deux images. Chacune de ces images est transformée en une matrice réduite ($F_i$ et $F_j$) de taille N*N où est extraite pour chaque cellule la valeur médiane vectorielle des pixels composant le bloc. Cette valeur médiane (ou pixel médian $P_x$) est représentée par un triplet de valeur (L,a*,b*) dans l'espace CIELab. On calcule les $N^2 = 256$ distances, basées sur la formule de différence de couleur (DeltaE76) du système colorimétrique ClE1976Lab où la distance entre deux points ($P_x$ et $P_y$) de cet espace est basée sur la distance euclidienne. Ensuite la moyenne de ces $N^2$ distances sert de métrique pour la comparaison des deux images.

3. Et finalement la troisième distance est simplement la **moyenne des distances pixel à pixel dans l'espace** $La^*b^*$(équation C.3)

$$d(m,n) = \frac{1}{R*C}\sum_{i=1}^{R}\sum_{j=1}^{C}\sqrt{(L_m(i,j) - L_n(i,j))^2 + (a^*_m(i,j) - a^*_n(i,j))^2 + (b^*_m(i,j) - b^*_n(i,j))^2} \tag{C.3}$$

   *Où $R$ = Le nombre de lignes et $C$ = Le nombre de colonnes dans l'image.*

### C.2.2 Distance entre clusters

Les distances $D(X,Y)$ entre deux clusters $X$ et $Y$ généralement utilisées sont le minimum, le maximum ou la moyenne pondérée des distances entre les singletons $(x_n)$ qui constituent les deux clusters.

– **Saut minimum “Single linkage" :**

$$D(X,Y) = min(d(x,y)) \text{ where } x \in X, y \in Y \tag{C.4}$$

– **Saut maximum “Complete linkage" :**

$$D(X,Y) = max(d(x,y)) \text{ where } x \in X, y \in Y \tag{C.5}$$

– **Saut moyen “Average linkage" :**

$$D(X,Y) = \frac{1}{\Omega(X) * \Omega(Y)} \sum_{i=1}^{\Omega(X)} \sum_{j=1}^{\Omega(Y)} d(x_i, y_j) \text{ where } x_i \in X, y_j \in Y \tag{C.6}$$

### C.2.3 Tests

On désire avoir un indicateur sur les performances de cette phase de classification. L'idée est d'obtenir N clusters lorsque l'utilisateur désire N images pour résumer le film. Dans cette optique il faudrait avoir une image de la vérité terrain par cluster. Par conséquent on définit l'***exactitude*** du système comme étant le nombre de cluster contenant *une et une seule image* de la vérité terrain (et qu'elle soit en plus différentes pour chaque cluster) sur le nombre de cluster (équation C.7).

$$Exactitude = \frac{\text{Nb cluster avec une et une seule image de la VT}}{\text{Nb cluster}} \tag{C.7}$$

On teste les différentes distances présentées dans ce paragraphe sur les deux meilleures configurations de l'algorithme à accumulation de différences c'est-à-dire : color naming 267 et DeltaE76 (voir figure C.2). Le tableau C.1 synthétise les résultats obtenus.

| *A&A méthode comparaison image :* ***DeltaE76****, Rappel :* ***0.89*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Méthode** | HC | HA | HS | BA | BS | BC | DS | DA | DC |
| **Exactitude** | 0.32 | 0.33 | 0.33 | 0.34 | 0.38 | 0.40 | 0.45 | 0.47 | 0.53 |

| *A&A méthode comparaison image :* ***Name267****, Rappel :* ***0.93*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Méthode** | HS | HA | BS | HC | BA | BC | DS | DA | DC |
| **Exactitude** | 0.30 | 0.35 | 0.35 | 0.39 | 0.40 | 0.41 | 0.45 | 0.51 | 0.53 |

TABLE C.1 – Comparaison des différentes méthodes de calcul de distances entre image et cluster *(méthode : H : Histogramme, B : Bloc, D : DeltaE76, C : Complete, S : Single, A : Average)*

Pour mieux évaluer les résultats présentés dans le tableau C.1 on propose de regrouper les résultats suivant la méthode utilisée pour comparer deux images entres elles et de faire la moyenne des valeurs correspondantes (voir tableau C.2). On fait de même suivant la méthode

utilisée pour comparer deux clusters entres eux (voir tableau C.3).

| **Méthode** | Block | Hist | DeltaE76 |
|---|---|---|---|
| **Exactitude** | 0.40 | 0.41 | 0.49 |

TABLE C.2 – Moyenne des résultats suivant la méthode de calcul des distances entre images

| **Méthode** | Single | Average | Complete |
|---|---|---|---|
| **Exactitude** | 0.38 | 0.40 | 0.43 |

TABLE C.3 – Moyenne des résultats suivant la méthode de calcul des distances entre clusters

### C.2.4 Discussions

On peut voir d'après les tableaux C.1 et C.3 que les méthodes de clustering *Complete* et *Average* donnent les meilleurs résultats par rapport à la méthode du saut minimum « Single linkage ». On peut noter qu'il y'a souvent des images de transition entre deux images différentes. Un désagrément majeur bien connu du saut minimum est l'effet de chaine qui met dans un même groupe deux objets éloignés lorsqu'il existe entre eux une suite de points peu éloignés les uns des autres. Ceci explique pourquoi la méthode du saut minimum donne de moins bons résultats. La mesure de similarité entre images basée sur la 3ième méthode (pixel à pixel) définie par l'équation C.3 donne les meilleurs résultats. Ceci s'explique par le fait que cette méthode considère chaque pixel de l'image et tend à donner des valeurs importantes de distances pour deux images même très similaires. Par contre elle est gourmande en temps de calcul. On peut noter que la méthode de réduction en bloc tend vers cette méthode lorsque la taille des blocs diminue. Empiriquement on constate que lorsqu'un bloc a une taille inférieure ou égale à une cinquantaine de pixels on obtient des résultats similaires mais avec un gain en temps de calcul non négligeable ($\approx 30$ sur des images de 800x600).

Annexe D

# Tests et résultats de l'analyse de texte

## D.1 Analyse syntaxique

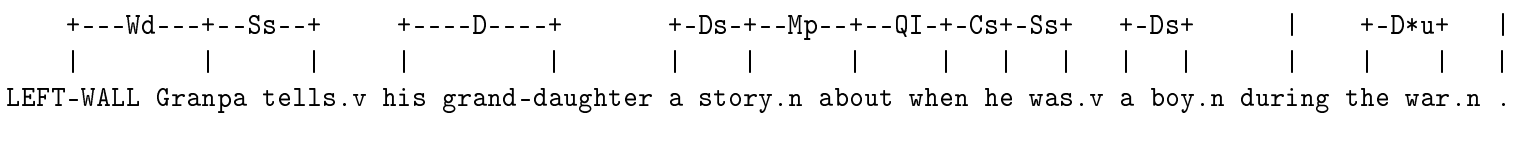
Voici quelques résultats complets de l'analyse syntaxique et de l'instanciation du scénario actanciel sur des phrases issues de synopsis de films d'animation. Voici ce qui est représenté sur les pages suivantes :

- La phrase issue du synopsis.
- Les différents « linkage » retournés par LG.
- Les tableaux statistiques des éléments (Sujet-Verbe-Objet-Adverbial) connectés aux verbes retrouvés.
- Le scénario actanciel instancié.

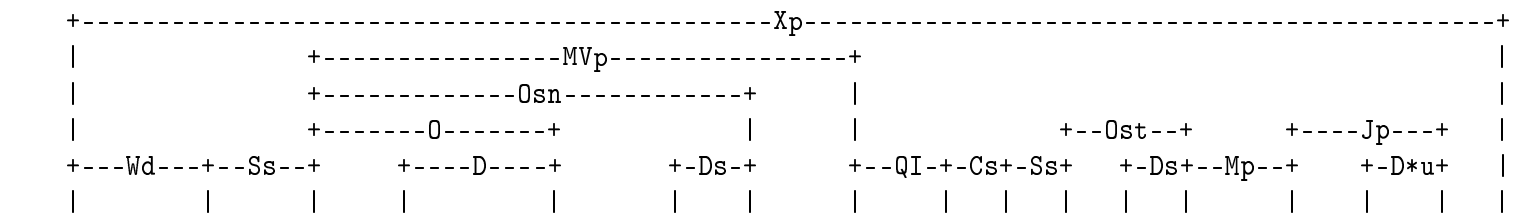
**Phrase du synopsis :** *Granpa tells his grand-daughter a story about when he was a boy during the war.*

**Les différents « linkage » retournés par LG :**

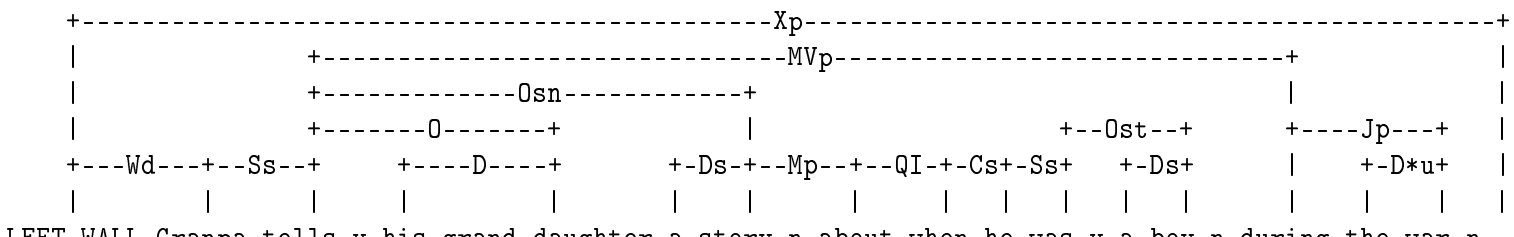
```
Linkage: 1/6 SubLinkage: 1/1

    +-----------------------------------------------Xp-----------------------------------------------+
    |               +----------------MVp----------------+                                            |
    |               +-------------Osn------------+      |             +------MVp-----+              |
    |               +-------O-------+            |      |             +--Ost--+      +----Jp---+    |
    +---Wd---+--Ss--+     +----D----+       +-Ds-+      +--QI-+-Cs+-Ss+   +-Ds+      |    +-D*u+   |
    |        |      |     |         |       |    |      |     |   |   |   |   |      |    |    |   |
LEFT-WALL Granpa tells.v his grand-daughter a story.n about when he was.v a boy.n during the war.n .


Linkage: 2/6 SubLinkage: 1/1

    +-----------------------------------------------Xp-----------------------------------------------+
    |               +-------------------------------MVp-------------------------------+             |
    |               +----------------MVp----------------+                              |             |
    |               +-------------Osn------------+      |                              |             |
    |               +-------O-------+            |      |             +--Ost--+      +----Jp---+    |
    +---Wd---+--Ss--+     +----D----+       +-Ds-+      +--QI-+-Cs+-Ss+   +-Ds+      |    +-D*u+   |
    |        |      |     |         |       |    |      |     |   |   |   |   |      |    |    |   |
LEFT-WALL Granpa tells.v his grand-daughter a story.n about when he was.v a boy.n during the war.n .


Linkage: 3/6 SubLinkage: 1/1

    +-----------------------------------------------Xp-----------------------------------------------+
    |               +-------------Osn------------+                     +------MVp-----+              |
    |               +-------O-------+            |                     +--Ost--+      +----Jp---+    |
```



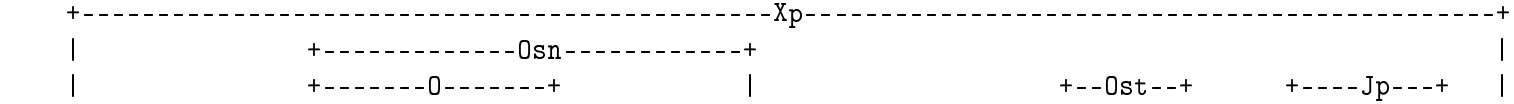

Linkage: 4/6 SubLinkage: 1/1




Linkage: 5/6 SubLinkage: 1/1




Linkage: 6/6 SubLinkage: 1/1



```
    +---Wd---+--Ss--+     +----D----+       +-Ds-+--Mp--+--QI-+-Cs+-Ss+   +-Ds+--Mp--+    +-D*u+   |
    |        |      |     |         |       |    |      |     |   |   |   |   |      |    |    |   |
LEFT-WALL Granpa tells.v his grand-daughter a story.n about when he was.v a boy.n during the war.n .
```

**La liste et les occurrences des verbes retrouvés**

```
Verb: 0 0 6 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0  Mean: 6 Std: 0
Verbs = {tells(2),was(10),}
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au premier verbe *(tells)* :**

```
S: 0 6 X 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  Mean: 6 Std: 0
V: 0 0 X 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  Mean: 0 Std: 0
O: 0 0 X 0 6 0 6 0 0 0 0 0 0 0 0 0 0 0 0  Mean: 6 Std: 0
A: 0 0 X 0 0 0 0 3 0 0 0 0 0 2 0 0 0 0 0  Mean: 2.5 Std: 0.7071

SVOA(1)

S = {Granpa }
V = {tells }
O = {grand-daughter, story }
A = {about, during }
```

**Recherche des groupes de mots connectés aux éléments précédents :**

```
SG: 00 06 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
OG: 00 00 00 06 06 00 00 00 00 00 00 00 00 00 00 00 00 00 00
OG: 00 00 00 00 00 06 06 00 00 00 00 00 00 00 00 00 00 00 00
AG: 00 00 00 00 00 00 00 06 06 06 06 06 06 00 00 00 00 00 00
AG: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 06 06 06 00 00

SVOAGroup(1)
```

```
Subj = { Granpa "Granpa" }
Verb = { tells }
Obj = { grand-daughter "his grand-daughter" }   { story "a story" }
Adv = { about "about when he was a boy" }   { during "during the war" }
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au deuxième verbe *(was)* :**

```
S: 0 0 0 0 0 0 0 0 0 6 X 0 0 0 0 0 0 0  Mean: 6 Std: 0
V: 0 0 0 0 0 0 0 0 0 0 X 0 6 0 0 0 0 0  Mean: 0 Std: 0
O: 0 0 0 0 0 0 0 0 0 0 X 0 6 0 0 0 0 0  Mean: 6 Std: 0
A: 0 0 0 0 0 0 0 0 0 0 X 0 0 2 0 0 0 0  Mean: 0 Std: 0

SVOA(2)

S = {he }
V = {was }
O = {boy }
A = {during }
```

**Recherche des groupes de mots connectés aux éléments précédents :**

```
SG: 00 00 00 00 00 00 00 00 00 06 00 00 00 00 00 00 00 00
OG: 00 00 00 00 00 00 00 00 00 00 00 06 06 00 00 00 00 00
AG: 00 02 02 02 02 02 02 02 02 00 00 00 00 06 06 06 00 00

SVOAGroup(2)

Subj = { he "he" }
Verb = { was }
Obj = { boy "a boy" }
Adv = { during "during the war" }
```

**Instanciation du Scenario Actanciel :**

```
**********
Actant:
**********
  |-Character = Granpa,
  |-Patient = grand-daughter "his grand-daughter", story "a story",


**********
Action:
**********
  |-Action = tell,


**********
Scene:
**********
  |-Locative =
  |-Temporal = war "during the war", <-- about when he was a boy -->
```

**Remarque :** <– *xxx* –> signifie que l'expression *xxx* constitue la scène.

**Phrase du synopsis :** *when she was a child, she was abused physically by a man.*

**Les différents « linkage » retournés par LG :**

```
Linkage: 1/1 SubLinkage: 1/1

    +-----------------------------------Xp-----------------------------------+
    +----------------Wd---------------+                                      |
    |       +----------CO*s----------+                                      |
    |       +----------Xc----------+  |                                      |
    |       |        +---Ost--+    |  |            +-------MVp------+--Js-+   |
    |       +-Cs+-Ss-+   +-Ds-+    |  +-Ss-+---Pv--+---MVa---+      | +-Ds+   |
    |       |   |    |   |    |    |  |    |       |         |      | |   |   |
LEFT-WALL when she was.v a child.n , she was.v abused.v physically by a man.n .
```

**La liste et les occurrences des verbes retrouvés**

```
VG: 00 00 00 01 00 00 00 00 01 01 00 00 00 00 00 00  Mean: 1 Std: 0
Verbs = {was(3),was(8),abused(9),}
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au premier verbe *(was)* :**

```
S: 00 00 01 X 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 1 Std: 0
V: 00 00 00 X 00 01 00 00 00 00 00 00 00 00 00 00  Mean: 0 Std: 0
O: 00 00 00 X 00 01 00 00 00 00 00 00 00 00 00 00  Mean: 1 Std: 0
A: 00 00 00 X 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 0 Std: 0

SVOA(1)

S = {she}
V = {was}
O = {child}
A = {}
```

**Recherche des groupes de mots connectés aux éléments précédents :**

```
SG: 00 00 01 00 00 00 00 00 00 00 00 00 00 00 00 00 00
OG: 00 00 00 00 01 01 00 00 00 00 00 00 00 00 00 00 00

SVOAGroup(1)
Subj = { she "she" }
Verb = { was }
Obj = { child "a child" }
Adv =
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au deuxième verbe *(was)* :**

```
S: 00 00 00 00 00 00 00 01 X 00 00 00 00 00 00 00  Mean: 1 Std: 0
V: 00 00 00 00 00 00 00 00 X 01 00 00 00 00 00 00  Mean: 1 Std: 0
O: 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00 00  Mean: 0 Std: 0
A: 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00 00  Mean: 0 Std: 0
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au troisième verbe *(abused)* :**

```
S: 00 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00  Mean: 0 Std: 0
V: 00 00 00 00 00 00 00 00 01 X 00 00 00 00 00 00  Mean: 1 Std: 0
O: 00 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00  Mean: 0 Std: 0
A: 00 00 00 00 00 00 00 00 00 X 01 01 00 00 00 00  Mean: 0 Std: 0
```

**Les deux verbes précédents sont liés ils constituent ainsi un groupe verbal**
**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au groupe verbal *(was abused)* :**

```
S: 00 00 00 00 00 00 00 01 X X 00 00 00 00 00 00  Mean: 1 Std: 0
V: 00 00 00 00 00 00 00 00 X X 00 00 00 00 00 00  Mean: 0 Std: 0
O: 00 00 00 00 00 00 00 00 X X 00 00 00 00 00 00  Mean: 0 Std: 0
A: 00 00 00 00 00 00 00 00 X X 01 01 00 00 00 00  Mean: 0 Std: 0
```

```
SVOA(2)

S = {she }
V = {was abused }
O = {}
A = {physically by}
```

**Recherche des groupes de mots connectés aux éléments précédents :**

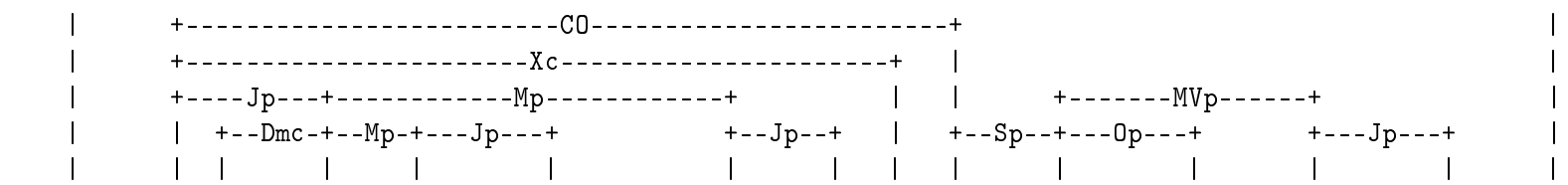
```
SG: 00 00 00 00 00 00 00 01 00 00 00 00 00 00 00 00
AG: 00 00 00 00 00 00 00 00 00 00 01 00 00 00 00 00
AG: 00 00 00 00 00 00 00 00 00 00 00 01 01 01 00 00

SVOAGroup(2)

Subj = { she "she" }
Verb = { was }
Obj = { child "a child" }
Adv =
```

**Instanciation du Scenario Actanciel :**
Aucun verbes d'action a été retrouvé !!!

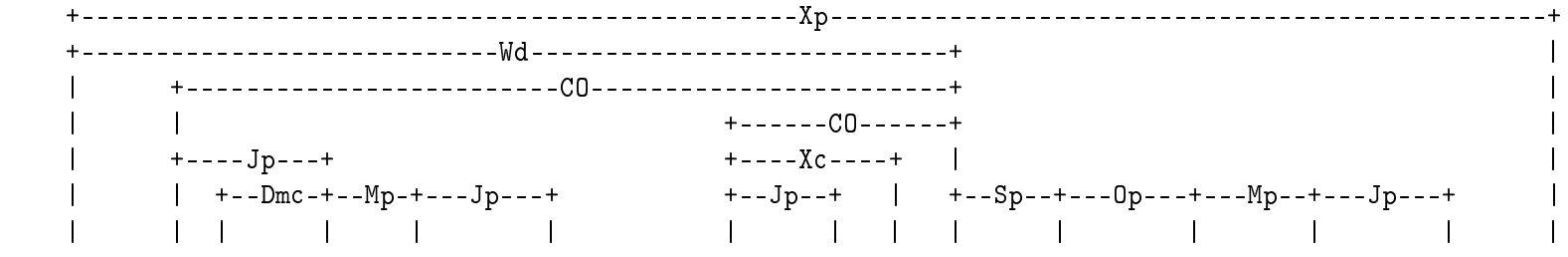
**Phrase du synopsis :** *On the themes of communication between men, three short stories about dialogues.*

**Les différents « linkage » retournés par LG :**

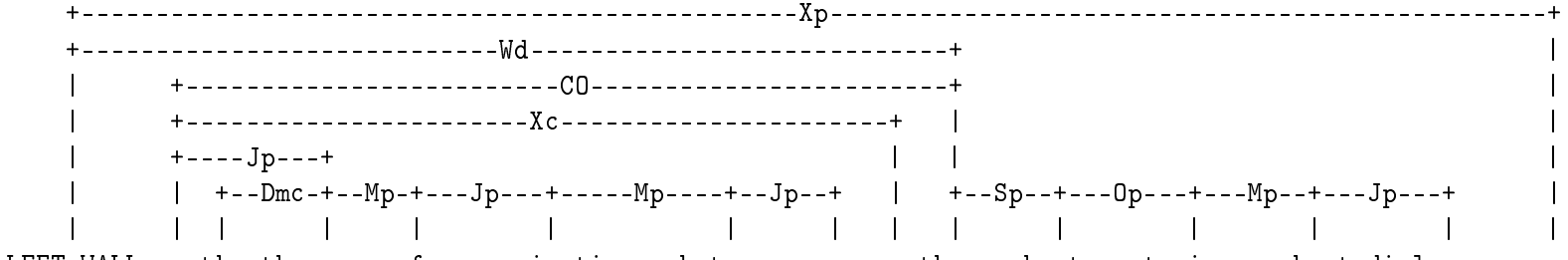
```
Linkage: 1/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
    |      +-------------------------CO-------------------------+                                       |
    |      |                                  +------CO------+                                        |
    |      +----Jp---+                        +----Xc----+   |      +-------MVp------+               |
    |      |  +--Dmc-+--Mp-+---Jp---+         +--Jp--+   |   +--Sp--+---Op---+       +---Jp---+      |
    |      |  |      |     |        |         |      |   |   |      |        |       |        |      |
LEFT-WALL on the themes.n of communication.n between men.n , three short.v stories.n about dialogues.n .


Linkage: 2/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
    |      +-------------------------CO-------------------------+                                       |
    |      +-----------------------Xc-----------------------+   |                                       |
    |      +----Jp---+                                     |   |      +-------MVp------+               |
    |      |  +--Dmc-+--Mp-+---Jp---+-----Mp----+--Jp--+   |   +--Sp--+---Op---+       +---Jp---+      |
    |      |  |      |     |        |           |      |   |   |      |        |       |        |      |
LEFT-WALL on the themes.n of communication.n between men.n , three short.v stories.n about dialogues.n .


Linkage: 3/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
```



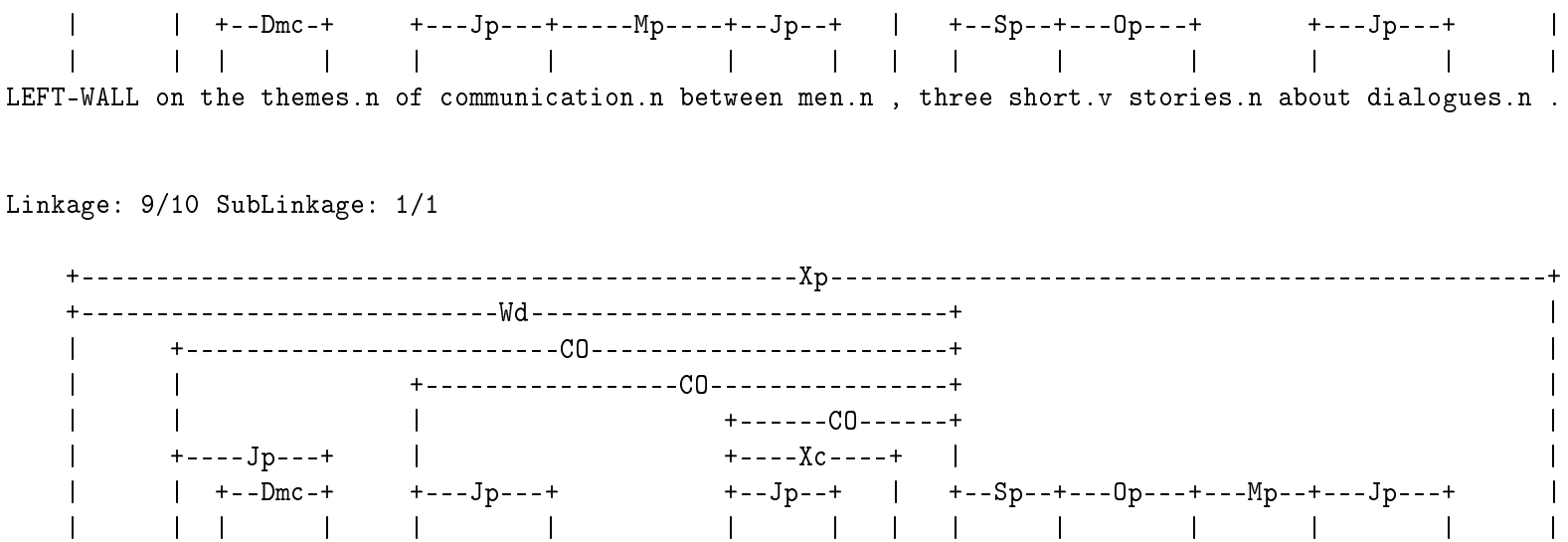

Linkage: 4/10 SubLinkage: 1/1




Linkage: 5/10 SubLinkage: 1/1



```
Linkage: 6/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
    |      +-------------------------CO-------------------------+                                       |
    |      +-----------------------Xc-----------------------+   |                                       |
    |      +----Jp---+-----------Mp-----------+          |   |                                       |
    |      |  +--Dmc-+--Mp-+---Jp---+          +--Jp--+   |   +--Sp--+---Op---+---Mp--+---Jp---+      |
    |      |  |      |     |        |          |      |   |   |      |        |       |        |      |
LEFT-WALL on the themes.n of communication.n between men.n , three short.v stories.n about dialogues.n .


Linkage: 7/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
    |      +-------------------------CO-------------------------+                                       |
    |      |               +-----------------CO----------------+                                       |
    |      |               |                  +------CO------+                                       |
    |      +----Jp---+     |                  +----Xc----+   |      +-------MVp------+               |
    |      |  +--Dmc-+     +---Jp---+          +--Jp--+   |   +--Sp--+---Op---+       +---Jp---+      |
    |      |  |      |     |        |          |      |   |   |      |        |       |        |      |
LEFT-WALL on the themes.n of communication.n between men.n , three short.v stories.n about dialogues.n .


Linkage: 8/10 SubLinkage: 1/1

    +-------------------------------------------------Xp-----------------------------------------------+
    +----------------------------Wd-----------------------------+                                       |
    |      +-------------------------CO-------------------------+                                       |
    |      |               +-----------------CO----------------+                                       |
    |      +----Jp---+     +---------------Xc--------------+   |      +-------MVp------+               |
```




**La liste et les occurrences des verbes retrouvés**

```
VG: 00 00 00 00 00 00 00 00 00 00 10 00 00 00 00 00  Mean: 10 Std: 0
Verbs = {short(10),}
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au premier verbe *(short)* :**

```
S: 00 00 00 00 00 00 00 00 00 10 X 00 00 00 00 00  Mean: 10 Std: 0
V: 00 00 00 00 00 00 00 00 00 00 X 00 00 00 00 00  Mean: 0 Std: 0
O: 00 00 00 00 00 00 00 00 00 00 X 10 00 00 00 00  Mean: 10 Std: 0
A: 00 00 00 00 00 00 00 00 00 00 X 00 05 00 00 00  Mean: 0 Std: 0

SVOA(1)

S = {three }
V = {short }
O = {stories }
A = {about }
```

**Recherche des groupes de mots connectés aux éléments précédents :**

```
SG: 00 00 00 00 00 00 00 00 00 10 00 00 00 00 00 00
OG: 00 00 00 00 00 00 00 00 00 00 00 10 00 00 00 00
AG: 00 00 00 00 00 00 00 00 00 00 00 00 10 10 00 00


SVOAGroup(1)

Subj = { three "three" }
Verb = { short }
Obj = { stories "stories" }
Adv = { about "about dialogues" }
```

**Instanciation du Scenario Actanciel :**

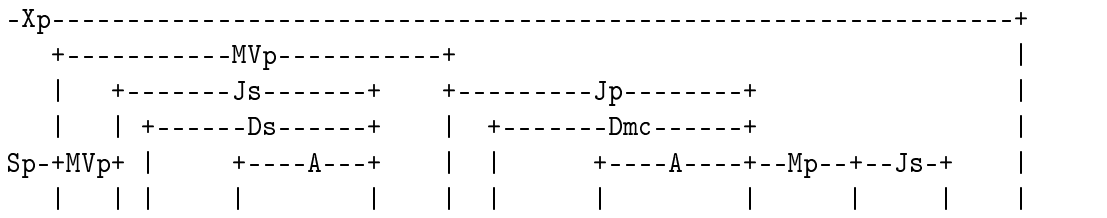
```
**********
Actant:
```

```
**********
  |-Character = three,
  |-Patient = stories,


**********
Action:
**********
  |-Action = short,


**********
Scene:
**********
  |-Locative =
  |-Temporal =
```

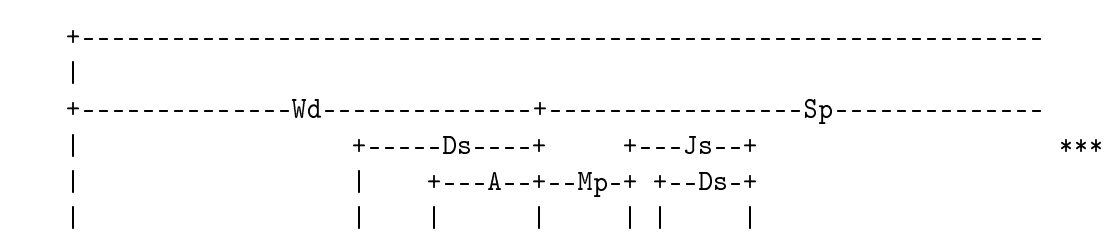
**Phrase du synopsis :** *Based on a true story of a father and son who go on a fishing trip in the untamed forests of Montreal.*

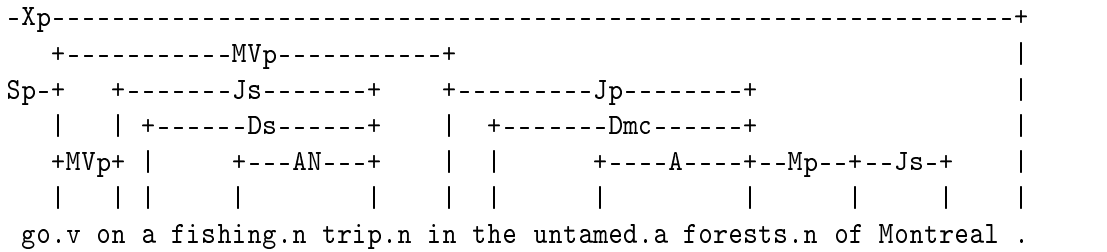
**Les différents « linkage » retournés par LG :**

```
Linkage: 1/54 SubLinkage: 1/2
    +-----------------------------------------------------------------
    |
    +--------------Wd--------------+-----------------Sp--------------
    |                  +-----Ds----+     +---Js--+                    ***
    |                  |    +---A--+--Mp-+ +--Ds-+
    |                  |    |      |     | |     |
LEFT-WALL [based] [on] a true.a story.n of a father.n and son.n [who]


-Xp-------------------------------------------------------------------+
   +-----------MVp-----------+                                        |
Sp-+   +-------Js-------+    +---------Jp--------+                   |
   |   | +------Ds------+    |  +-------Dmc------+                   |
   +MVp+ |     +----A---+    |  |      +----A----+--Mp--+--Js-+    |
   |   | |     |        |    |  |      |         |      |     |    |
 go.v on a fishing.g trip.n in the untamed.a forests.n of Montreal .

SubLinkage: 2/2


    +-----------------------------------------------------------------
    |
    +--------------------------Wd--------------------------+         ***
    |                  +-----------------Ds-----------------+
    |                  |    +---------------A---------------+-----Sp-
    |                  |    |                               |
LEFT-WALL [based] [on] a true.a story.n of a father.n and son.n [who]
```

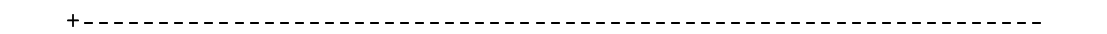



Linkage: 2/54 SubLinkage: 1/2

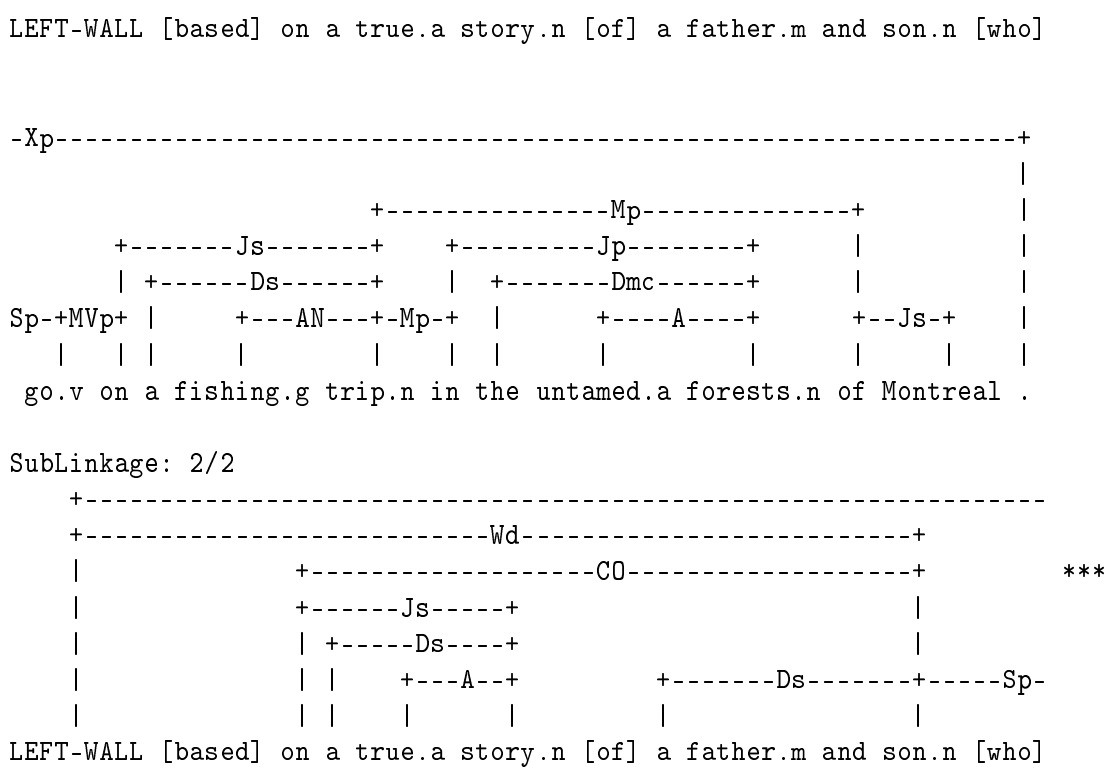



SubLinkage: 2/2



```
    |
    +---------------------------Wd--------------------------+         ***
    |                   +-----------------Ds-----------------+
    |                   |    +---------------A---------------+-----Sp-
    |                   |    |                               |
LEFT-WALL [based] [on] a true.a story.n of a father.n and son.n [who]


-Xp-------------------------------------------------------------------+
   +-----------MVp-----------+                                        |
   |   +-------Js-------+    +---------Jp--------+                   |
   |   | +------Ds------+    |  +-------Dmc------+                   |
Sp-+MVp+ |     +---AN---+    |  |      +----A----+--Mp--+--Js-+    |
   |   | |     |        |    |  |      |         |      |     |    |
 go.v on a fishing.n trip.n in the untamed.a forests.n of Montreal .


*
*
*



Linkage: 54/54 SubLinkage: 1/2

    +-----------------------------------------------------------------
    +---------------------Wd---------------------+
    |              +--------------CO-------------+                    ***
    |              +------Js-----+               |
    |              | +-----Ds----+               |
    |              | |    +---A--+         +--Ds-+----------Sp-------
    |              | |    |      |         |     |
```


LEFT-WALL [based] on a true.a story.n [of] a father.m and son.n [who]
-Xp
Mp
Js
Jp
Ds
Dmc
Sp-+MVp+
AN
Mp-+
A
Js-+
go.v on a fishing.g trip.n in the untamed.a forests.n of Montreal .
SubLinkage: 2/2
Wd
CO
***
Js
Ds
A
Ds
Sp-
LEFT-WALL [based] on a true.a story.n [of] a father.m and son.n [who]
-Xp
Mp
Js
Jp
Ds
Dmc
Sp-+MVp+
AN
Mp-+
A
Js-+
go.v on a fishing.g trip.n in the untamed.a forests.n of Montreal .

**La liste et les occurrences des verbes retrouvés**

```
Verbs: 00 00 00 00 00 00 00 00 00 00 00 00 54 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 54 Std: 0
Verbs = {go(12),}
```

**Recherche des éléments Sujet-Verbe-Objet-Adverbial connectés au premier verbe *(tells)* :**

```
S: 00 00 00 00 00 36 00 00 18 00 54 00 X 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 36 Std: 12.7279
V: 00 00 00 00 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 0 Std: 0
O: 00 00 00 00 00 00 00 00 00 00 00 00 X 00 00 00 00 00 00 00 00 00 00 00 00  Mean: 0 Std: 0
A: 00 00 00 00 00 00 00 00 00 00 00 00 X 108 00 00 00 36 00 00 00 00 00 00 00  Mean: 72 Std: 36

SVOA(1)

S = {story father son }
V = {go }
O = {}
A = {on }
```

**Recherche des groupes de mots connectés aux éléments précédents :**

```
SG: 00 00 36 72 72 108 36 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
SG: 00 00 00 00 00 00 36 54 108 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
SG: 00 00 00 00 00 00 00 00 00 00 108 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
AG: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 108 108 108 108 72 72 72 72 72 72 00 00

SVOAGroup(1)

Subj = { story "a true story" }   { father "a father" }   { son "son" }
Verb = { go }
Obj =
Adv = { on "on a fishing trip in the untamed forests of Montreal" }
```

**Instanciation du Scenario Actanciel :**

```
**********
Actant:
**********
  |-Character = father " a father ", son, story " a true story ",
  |-Patient =


**********
Action:
**********
  |-Action = go,


**********
Scene:
**********
  |-Locative = Montreal " on a fishing trip in the untamed forests of Montreal ",
  |            forests " on a fishing trip in the untamed forests of Montreal ",
  |-Temporal = fishing " on a fishing trip in the untamed forests of Montreal ",
  |            trip " on a fishing trip in the untamed forests of Montreal ",
```

## D.2 Classification supervisée des synopsis suivant les genres des films d'animation

Nous présentons ici une approche de classification supervisée des synopsis suivant les genres d'animation par analyse de textes. Le but de cette approche est d'obtenir automatiquement le genre d'un film à partir de l'analyse de son synopsis et de son vocabulaire. Notre approche se décompose en plusieurs phases décrites ici :

– La première étape est la phase d'apprentissage supervisé qui consiste à isoler les termes spécifiques de chacune des catégories du genre déclaré. Pour cela on calcul pour l'ensemble des termes $m$ du corpus et pour chacune des catégories du genre $\lambda$ (avec $\lambda \in$ [Artistique, Aventure, Publicitaire, Fantastique, Documentaire, Dramatique, Érotique, Expérimental, Humoristique, Musical, Satire, Policier, Politique, Western]) leur indice de spécificité (voir la définition de cet indice dans le §4.5.1 et équation 4.5). Cette opération est faite sur un corpus d'apprentissage constitué de 5804 synopsis non lemmatisés et nettoyés des mots outils. Pour chacune des catégories du genre déclaré nous constituons une référence lexicale (voir figure D.1) qui est composée d'une liste de termes associés à un indice de spécificité. D'après l'équation 4.5 l'indice de spécificité est dans l'intervalle $I^{\lambda}_{spe} \in [0; +\infty[$. Ainsi nous décidons de répartir cette mesure (voir équation D.1)

$$\widetilde{I^{\lambda}_{spe}}(m) = \begin{cases} -\frac{1}{I^{\lambda}_{spe}(m)} & \text{si } I^{\lambda}_{spe}(m) < 1 \\ I^{\lambda}_{spe}(m) & \text{sinon} \end{cases} \tag{D.1}$$

Avec $\widetilde{I^{\lambda}_{spe}}(m)$ l'indice de spécificité répartie pour le terme $m$ de la catégorie $\lambda$ et $I^{\lambda}_{spe}(m)$ l'indice de spécificité du terme $m$ de la catégorie $\lambda$. Finalement $\widetilde{I^{\lambda}_{spe}}(m) \in ]-\infty; +\infty[$.

| Artistique | Spécificité | Aventure | Spécificité | Publicitaire | Spécificité | Fantastique | Spécificité |
|---|---|---|---|---|---|---|---|
| +conçu | 37.22 | +vont | 5.53 | +publicitaire | 45.05 | +interroge | 5.99 |
| +imaginaire | 37.22 | +vent | 5.38 | +Publicité | 43.9 | +désert | 5.39 |
| +tableaux | 37.22 | +neige | 4.94 | +Spot | 32.76 | +légende | 5.39 |
| +Cirque | 31.01 | +triste | 4.61 | +robots | 19.66 | +perdu | 5.39 |
| +réaliser | 31.01 | +route | 4.32 | +vêtements | 19.66 | +portes | 5.39 |
| +séquences | 31.01 | +amitié | 4.15 | +Message | 10.92 | +contes | 5.39 |
| +imaginaires | 31.01 | +boule | 4.15 | +destiné | 10.92 | +princesse | 4.99 |
| +passent | 31.01 | +causé | 4.15 | +détruit | 9.83 | +paysage | 4.49 |
| +rythmes | 31.01 | +assis | 4.15 | +nucléaire | 9.83 | +promenade | 4.49 |
| +gens | 26.58 | +cosmos | 4.15 | +joyeux | 9.83 | +quitte | 4.49 |
| +présenté | 26.58 | +destin | 4.15 | +urbain | 9.83 | +bois | 4.49 |
| +théâtre | 26.58 | +oeuf | 4.15 | +édition | 9.83 | +population | 4.49 |
| +concert | 26.58 | +solidarité | 4.15 | +Frederic | 9.83 | +conte | 4.19 |
| +lune | 26.58 | +Jules | 4.15 | +tellement | 9.83 | +étranges | 4.14 |
| +unis | 23.26 | +avion | 3.95 | +joue | 8.94 | +pierre | 3.99 |
| +leçon | 23.26 | +loup | 3.95 | +société | 8.78 | +chasse | 3.99 |
| +gestes | 23.26 | +amis | 3.77 | +peintre | 8.19 | +matière | 3.85 |
| +princesse | 20.68 | +Aventures | 3.71 | +chante | 8.19 | +naît | 3.85 |
| +fête | 18.61 | +bonhomme | 3.46 | +compagnie | 8.19 | +lune | 3.85 |

FIGURE D.1 – Extrait des références lexicales (terme + indice de spécificité répartie) pour les catégories : Artistique, Aventure, Publicitaire et Fantastique

– Dans la deuxième étape nous testons le pouvoir discriminant des termes spécifiques appris précédemment. Ainsi, nous déterminons pour chaque synopsis son genre le plus probable que nous comparons in fine au genre déclaré. Pour chaque synopsis $S$ et pour chacune des catégories $\lambda$ des genres, nous calculons un indice $I_\lambda(S)$ qui traduit l'appartenance du synopsis $S$ à la catégorie $\lambda$. Soit pour le synopsis $S$ le calcul de 14 indices $(I_{Artistique}(S), I_{Aventure}(S), ..., I_{Western}(S))$ liés à chacune des catégories du genre. Le calcul de ces indices se fait à partir des références lexicales $\Omega_\lambda$ (qui ont été constituées auparavant durant la phase d'apprentissage) et correspond à la somme des indices de spécificité répartie des termes appartenant au synopsis et à la référence lexicale $\Omega_\lambda$ (voir l'équation D.2) où le synopsis $S$ et la référence lexicale $\Omega_\lambda$ sont vus comme des ensembles de termes $m$.

$$I_\lambda(S) = \sum_{m \in S \cap \Omega_\lambda} \widetilde{I^\lambda_{spe}}(m) \tag{D.2}$$

Le calcul de ces indices $I_\lambda(S)$ est réalisé sur une base de test constituée d'un peu plus de 18150 synopsis contenant la base d'apprentissage.
Nous partons de l'hypothèse que les références lexicales apprises ainsi que le calcul de l'indice $I_\lambda(S)$ permettent de discriminer les catégories du genre. Cependant il nous apparaît intéressant de tester la véracité de cette hypothèse. Pour cela nous calculons pour chacune des catégories $\lambda$, ses paramètres statistiques $\mu_{I_\lambda}(S \in \lambda)$, $\sigma_{I_\lambda}(S \in \lambda)$ et $\mu_{I_\lambda}(S \notin \lambda)$, $\sigma_{I_\lambda}(S \notin \lambda)$ correspondant respectivement à la moyenne et à l'écart type des indices $I_\lambda(S)$ des synopsis $S$ dont le genre déclaré appartient à la catégorie $\lambda$ et des synopsis $S$ dont le genre déclaré n'appartient pas à la catégorie $\lambda$. Par conséquent l'indice $I_\lambda$ permet de discriminer les genres si $\mu_{I_\lambda}(S \in \lambda) \neq \mu_{I_\lambda}(S \notin \lambda)$. Nous proposons de comparer ces deux moyennes et nous posons $\Delta_\mu = \mu_{I_\lambda}(S \in \lambda) - \mu_{I_\lambda}(S \notin \lambda)$. La division de $\Delta_\mu$ par son écart type (donné par l'équation D.3) suit une loi normale centrée réduite de moyenne 0 et d'écart type 1.

$$\sigma_{\Delta_\mu} = \sqrt{\frac{\sigma^2_{I_\lambda(S \in \lambda)}}{card(S \in \lambda)} + \frac{\sigma^2_{I_\lambda(S \notin \lambda)}}{card(S \notin \lambda)}} \tag{D.3}$$

$$Z = \frac{|\Delta_\mu|}{\sigma_{\Delta_\mu}} \tag{D.4}$$

Finalement le test de Z ou de l'écart réduit (voir équation D.4) permet d'accepter l'hypothèse avec un niveau de confiance $\alpha$ si $\alpha < \Phi(Z)$ où $\Phi$ désigne la fonction de répartition de la loi normale centrée réduite. Les résultats sont présentés sur la figure D.2.

| | Nombre oui | Moyenne oui | Ecart-type oui | Nombre non | Moyenne non | Ecart-type non | Différence moyennes | Ecart-type moyenne | Test Loi normale | Résultat test Z |
|---|---|---|---|---|---|---|---|---|---|---|
| Artistique | 2430 | 7.667502 | 12.600984 | 15725 | 7.174363 | 11.101372 | 0.4931 | 0.2705 | 0.965843268366669 | Différence significative |
| Aventure | 990 | 8.459543 | 7.3541243 | 17165 | 3.475428 | 6.2547652 | 4.9841 | 0.2386 | 1.000000000000000 | Différence très significative |
| Publicitaire | 178 | 38.71347 | 35.984942 | 17977 | 3.729975 | 11.35469 | 34.9835 | 2.6985 | 1.000000000000000 | Différence très significative |
| Fantastique | 1184 | 4.073555 | 5.2696972 | 16971 | 2.196258 | 4.5890361 | 1.8773 | 0.1571 | 1.000000000000000 | Différence très significative |
| Documentaire | 865 | 6.122121 | 7.8170884 | 17290 | 5.247373 | 12.405446 | 0.8747 | 0.2820 | 0.999037422796264 | Différence très significative |
| Dramatique | 1044 | 5.082446 | 4.9884063 | 17111 | 3.757677 | 4.6935128 | 1.3248 | 0.1585 | 1.000000000000000 | Différence très significative |
| Erotique | 187 | 9.219118 | 11.391807 | 17968 | 4.598343 | 5.9809933 | 4.6208 | 0.8342 | 0.999999984778268 | Différence très significative |
| Expérimental | 926 | 12.96172 | 22.682058 | 17229 | 4.001163 | 7.3702962 | 8.9606 | 0.7475 | 1.000000000000000 | Différence très significative |
| Humoristique | 3732 | 1.469116 | 3.6185699 | 14423 | 1.028862 | 3.5757105 | 0.4403 | 0.0663 | 0.999999999984402 | Différence très significative |
| Musical | 960 | 8.715102 | 10.600917 | 17195 | 2.350903 | 5.547837 | 6.3642 | 0.3447 | 1.000000000000000 | Différence très significative |
| Satire | 633 | 1.686896 | 4.163506 | 17522 | -0.189899 | 3.8616304 | 1.8768 | 0.1680 | 1.000000000000000 | Différence très significative |
| Policier | 28 | 44.34143 | 98.567343 | 18127 | 10.21438 | 20.311227 | 34.1271 | 18.6281 | 0.966525875089578 | Différence significative |
| Politique | 254 | 7.535591 | 18.810752 | 17901 | 2.454777 | 6.9587623 | 5.0808 | 1.1814 | 0.999991480775399 | Différence très significative |
| Western | 21 | 28.17238 | 76.255439 | 18134 | 2.513193 | 9.2197906 | 25.6592 | 16.6404 | 0.938460476762311 | Différence non significative |

FIGURE D.2 – Comparaison des moyennes des indices $I_\lambda(S \in \lambda)$ et $I_\lambda(S \notin \lambda)$ par le test de Z. *"Nombre oui" correspond à* $card(S \in \lambda)$, *"Moyenne oui" correspond à* $\mu_{I_\lambda}(S \in \lambda)$, *"Ecart-type oui" correspond à* $\sigma_{I_\lambda}(S \in \lambda)$, *"Nombre non" correspond à* $card(S \notin \lambda)$, *"Moyenne non" correspond à* $\mu_{I_\lambda}(S \notin \lambda)$, *"Ecart-type non" correspond à* $\sigma_{I_\lambda}(S \notin \lambda)$, *"Différence moyennes" correspond à* $\Delta_\mu$, *"Ecart-type moyenne" correspond à* $\sigma_{\Delta_\mu}$, *"Test Loi normale" correspond à* $\Phi(Z)$, *"Résultat test Z" correspond à* $\alpha < \Phi(Z)$ *avec* $\Phi(Z) < 0.95$ *(Différence non significative),* $0.95 \leqslant \Phi(Z) < 0.99$ *(Différence significative),* $0.99 \leqslant \Phi(Z)$ *(Différence très significative)*

– La troisième étape consiste à attribuer à chaque synopsis $S$ le genre le plus probable suivant les indices d'appartenance aux catégories $I_\lambda(S)$. Pour permettre une comparaison entre ces indices il est au préalable nécessaire de les **normaliser**. Pour cela nous considérons l'ensemble des indices d'appartenance à la catégorie $\lambda$ comme des répartitions statistiques définies par une moyenne $\mu_\lambda$ et un écart-type $\sigma_\lambda$. Cette répartition peut être transformée en une autre distribution statistique (voir équation D.5) qui a pour moyenne 0 et pour écart-type 1. La répartition de ce nouvel indice d'appartenance aux catégories $\overline{I_\lambda}(S)$ est dite « centrée réduite ». L'intérêt de standardiser cette variable est de pouvoir la comparer aux autres variables numériques.

$$\overline{I_\lambda}(S) = \frac{I_\lambda(S) - \mu_\lambda}{\sigma_\lambda} \tag{D.5}$$

Finalement la comparaison entre ces nouveaux indices d'appartenance aux catégories est trivial et nous attribuons au synopsis $S$ la catégorie $\lambda$ pour laquelle l'indice $\overline{I_\lambda}(S)$ est maximum (voir équation D.6).

$$Genre(S) = \arg\max_\lambda \overline{I_\lambda}(S) \tag{D.6}$$

– La quatrième étape consiste à comparer les valeurs de genre calculées à celles déclarées (prises comme vérité terrain). Nous utilisons les mesures de précision et de rappel pour quantifier la qualité de la classification.

$$\text{précision} = \frac{\text{synopsis de la catégorie } \lambda \cap \text{synopsis attribué à la catégorie } \lambda}{\text{synopsis attribué à la catégorie } \lambda} \tag{D.7}$$

$$\text{rappel} = \frac{\text{synopsis de la catégorie } \lambda \cap \text{synopsis attribué à la catégorie } \lambda}{\text{synopsis de la catégorie } \lambda} \tag{D.8}$$

Les résultats sont présentés sur la figure D.3.

| | Nombre oui | Nombre non | Résultat test Z | Total | Attribués au genre | Correctement attribué au genre | Rappel | Précision |
|---|---|---|---|---|---|---|---|---|
| Publicitaire | 178 | 17977 | Différence très significative | 18155 | 307 | 97 | 54% | 32% |
| Artistique | 2430 | 15725 | Différence significative | 18155 | 4329 | 1290 | 53% | 30% |
| Musical | 960 | 17195 | Différence très significative | 18155 | 493 | 202 | 21% | 41% |
| Expérimental | 926 | 17229 | Différence très significative | 18155 | 364 | 148 | 16% | 41% |
| Aventure | 990 | 17165 | Différence très significative | 18155 | 550 | 141 | 14% | 26% |
| Fantastique | 1184 | 16971 | Différence très significative | 18155 | 551 | 99 | 8% | 18% |
| Satire | 633 | 17522 | Différence très significative | 18155 | 498 | 62 | 10% | 12% |
| Humoristique | 3732 | 14423 | Différence très significative | 18155 | 513 | 232 | 6% | 45% |
| Politique | 254 | 17901 | Différence très significative | 18155 | 461 | 35 | 14% | 8% |
| Dramatique | 1044 | 17111 | Différence très significative | 18155 | 542 | 77 | 7% | 14% |
| Erotique | 187 | 17968 | Différence très significative | 18155 | 622 | 18 | 10% | 3% |
| Policier | 28 | 18127 | Différence significative | 18155 | 522 | 6 | 21% | 1% |
| Western | 21 | 18134 | Différence non significative | 18155 | 445 | 4 | 19% | 1% |
| Documentaire | 865 | 17290 | Différence très significative | 18155 | 125 | 6 | 1% | 5% |

FIGURE D.3 – Résultats de la classification automatique des synopsis

Les résultats de cette classification automatique sont présentés sur la figure D.3. On remarque que les résultats sont disparates et que les résultats dépassent rarement 50% de precision et/ou de rappel. Cependant on remarque que cette approche donne les meilleurs résultats pour les catégories *publicitaire* et *artistique* que l'on retrouve une fois sur deux. Cela vient de l'utilisation d'un vocabulaire spécialisé ou tout au moins spécifique à ces catégories particulières comme les termes "publicitaire", "publicité", "spot" pour le genre **publicitaire** ou les termes "imaginaire", "tableaux", "cirque" pour le genre **artistique**.

## D.3 Analyse thématique

### D.3.1 Thématique du Drame

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| jimmy | 8,39 | respirer | 5,6 | détester | 5,6 | meurtrir | 5,6 |
| julie | 8,39 | revivre | 5,6 | détrôner | 5,6 | minutieux | 5,6 |
| alcoolisme | 8,39 | typique | 5,6 | disponible | 5,6 | mitrailler | 5,6 |
| responsable | 7,46 | accoutumance | 5,6 | disséquer | 5,6 | mont | 5,6 |
| agressivité | 7,46 | affolement | 5,6 | dormeur | 5,6 | morne | 5,6 |
| baignoire | 7,46 | agent_secret | 5,6 | égypte | 5,6 | mortalité | 5,6 |
| becky | 7,46 | alcoolisé | 5,6 | embarrassant | 5,6 | moufle | 5,6 |
| comédien | 7,46 | allégorique | 5,6 | emiliana | 5,6 | naïveté | 5,6 |
| dante | 7,46 | ambigu | 5,6 | emprise | 5,6 | nationalité | 5,6 |
| destination | 7,46 | anatomiste | 5,6 | en_général | 5,6 | nelson | 5,6 |
| encadrer | 7,46 | antihéros | 5,6 | envieux | 5,6 | noyau | 5,6 |
| épidémie | 7,46 | anton | 5,6 | étroitesse | 5,6 | obéissant | 5,6 |
| handicaper | 7,46 | aoét | 5,6 | fidélité | 5,6 | obstiner | 5,6 |
| mineur | 7,46 | arabe | 5,6 | finlandais | 5,6 | palier | 5,6 |
| pingu | 7,46 | assaillir | 5,6 | flea | 5,6 | peiner | 5,6 |
| port | 7,46 | assoiffé | 5,6 | french | 5,6 | pelucher | 5,6 |
| porter_un_toast | 7,46 | assumer | 5,6 | gelé | 5,6 | peste | 5,6 |
| prévenir | 7,46 | astronome | 5,6 | genou | 5,6 | pétrifié | 5,6 |
| prostituée | 7,46 | attente | 5,6 | gong | 5,6 | poing | 5,6 |
| racisme | 7,46 | au_hasard | 5,6 | goutte_de_sang | 5,6 | point_de_départ | 5,6 |
| solliciter | 7,46 | au_possible | 5,6 | haie | 5,6 | pour_autant | 5,6 |
| soulever | 7,46 | bac | 5,6 | haillon | 5,6 | projectionniste | 5,6 |
| succomber | 7,46 | banalement | 5,6 | hanté | 5,6 | promoteur | 5,6 |
| annie | 7,46 | blessé | 5,6 | héroïsme | 5,6 | radeau | 5,6 |
| atypique | 7,46 | calciner | 5,6 | home | 5,6 | récemment | 5,6 |
| bouffon | 7,46 | cargo | 5,6 | hongrie | 5,6 | réciter | 5,6 |
| complexité | 7,46 | cendre | 5,6 | hugo | 5,6 | recourir | 5,6 |
| enlacer | 7,46 | chambouler | 5,6 | humble | 5,6 | recycler | 5,6 |
| fuite | 7,46 | charité | 5,6 | iconoclaste | 5,6 | redescendre | 5,6 |
| néant | 7,46 | chef_d | 5,6 | ignorant | 5,6 | redonner | 5,6 |
| pot_de_fleurs | 7,46 | chevalerie | 5,6 | impétueux | 5,6 | regret | 5,6 |
| résignation | 7,46 | chihuahua | 5,6 | importun | 5,6 | réincarner | 5,6 |
| corée | 6,72 | christian | 5,6 | inachevé | 5,6 | réveillon | 5,6 |
| sida | 6,72 | coïncidence | 5,6 | inceste | 5,6 | revendre | 5,6 |
| affection | 6,72 | collectionneur | 5,6 | incompris | 5,6 | sceller | 5,6 |
| franz | 5,6 | commerce | 5,6 | inconcevable | 5,6 | sénile | 5,6 |
| cercle | 5,6 | condenser | 5,6 | interprétant | 5,6 | soucier | 5,6 |
| fixer | 5,6 | confectionner | 5,6 | intimidation | 5,6 | standardisation | 5,6 |
| punir | 5,6 | contester | 5,6 | intrigant | 5,6 | stanley | 5,6 |
| anonyme | 5,6 | contradictoire | 5,6 | intrusion | 5,6 | strauss | 5,6 |
| idylle | 5,6 | contrée | 5,6 | irlande | 5,6 | supériorité | 5,6 |
| idyllique | 5,6 | convoitise | 5,6 | kabal | 5,6 | superstition | 5,6 |
| irlandais | 5,6 | country | 5,6 | koala | 5,6 | surpasser | 5,6 |
| remémorer | 5,6 | croissance | 5,6 | lampadaire | 5,6 | surpopulation | 5,6 |
| veille | 5,6 | cruauté | 5,6 | liaison | 5,6 | taciturne | 5,6 |
| affecter | 5,6 | cruellement | 5,6 | lieue | 5,6 | teinté | 5,6 |
| basculer | 5,6 | dansant | 5,6 | lugubre | 5,6 | tentaculaire | 5,6 |
| faire_nuit | 5,6 | découler | 5,6 | madrid | 5,6 | tout_de_suite | 5,6 |
| faire_plaisir | 5,6 | dégénérer | 5,6 | malveillant | 5,6 | toxique | 5,6 |
| infernal | 5,6 | délit | 5,6 | manager | 5,6 | trente | 5,6 |
| menaçant | 5,6 | désaffecté | 5,6 | maréchal | 5,6 | trouver_la_mort | 5,6 |
| purger | 5,6 | désireux | 5,6 | marine | 5,6 | tuer_le_temps | |

FIGURE D.4 – Liste des 100 premiers mots spécifiques du genre drame avec l'indice de spécificité

| | | |
|---|---|---|
| à_tout_jamais | envahir | séparer |
| abuser | errer | sida |
| accident | fuir | sinistre |
| agé | fusil | soldat |
| agressivité | fusiller | solitude |
| alcoolisme | guerre | sombrer |
| aliénation | horrible | subir |
| angoisse | impitoyable | succomber |
| arme | incident | suicider |
| attaque | infernal | tempête |
| battre | macabre | terrible |
| camp | mal | tragédie |
| cauchemar | malade | tragique |
| chagrin | malheureux | triste |
| chaos | mauvais | tuer |
| combat | méchant | vieil |
| combattre | menacer | vieillesse |
| conflit | meurtre | vieux |
| criminel | militaire | violence |
| cruel | monotone | violent |
| danger | morne | violer |
| dangereux | mort | |
| déchaîner | mourir | |
| démon | mystérieux | |
| déserter | mystique | |
| détruire | pauvreté | |
| dévaster | péché | |
| dévorer | peur | |
| diabolique | pire | |
| difficile | pleurer | |
| disparaître | pourri | |
| dispute | prisonnier | |
| drame | prostituée | |
| drogue | punir | |
| échapper | punition | |
| écraser | purger | |
| emprisonner | racisme | |
| Enfer | sacrifier | |
| enfermer | sang | |
| enfuir | séparation | |

FIGURE D.5 – Liste des germes pour la constitution du dictionnaire thématique du drame

### D.3.2 Thématique du Policier

Comme pour la thématique du Drame nous avons créé un dictionnaire thématique du Policier qui contient un peu plus de 150 termes dont la figure D.7 représente les 120 premiers termes. Pour vérifier le pouvoir discriminant de la mesure de l'intensité thématique du policier nous calculons pour chaque synopsis son intensité thématique puis nous calculons la moyenne des intensités thématiques pour chacune des catégories du genre déclaré sans tenir compte des non-réponses. On voit sur la figure D.6 que cette intensité thématique est importante et significative. Les noms des critères discriminants sont encadrés et correspondent à des moyennes significativement différentes de l'ensemble de l'échantillon au risque de 95% (test de student) dans le cas du genre **Policier**.

| Genre | V7Filtre_DictPolicier_I |
|---|---|
| __litterature | 0.00 (0.00 ) |
| _argent | 0.00 (0.00 ) |
| _educatif | 2.51 (0.36 ) |
| _provocateur | 3.51 (0.79 ) |
| Aventure | 2.52 (0.63 ) |
| Comédie | 0.55 (0.16 ) |
| Comédie_musicale | 0.89 (0.13 ) |
| Conte | 1.87 (0.39 ) |
| Documentaire | 1.20 (0.17 ) |
| Drame | 2.17 (0.46 ) |
| Epopée | 0.00 (0.00 ) |
| Erotique | 4.44 (0.94 ) |
| Expérimental | 2.77 (0.44 ) |
| Fantastique | 2.36 (0.47 ) |
| Humour | 3.14 (0.57 ) |
| Humour_noir | 0.00 (0.00 ) |
| Musical | 1.61 (0.23 ) |
| Policier | 12.33 (10.49 ) |
| Promotion | 0.00 (0.00 ) |
| Propagande | 3.62 (0.89 ) |
| Publicité | 2.37 (0.28 ) |
| Satire | 2.63 (0.52 ) |
| Science-fiction | 3.63 (1.60 ) |
| TOTAL | 2.91 (0.54 ) |

FIGURE D.6 – Moyenne des intensités thématiques du policier et écart-types entre parenthèses en fonction du genre déclaré

| | | |
|---|---|---|
| à_bout_portant | dérober | inculpation |
| agent | détective | inculper |
| agent_spécial | détention | indice |
| agence_spéciale | dissection | inspecteur |
| arme_à_feu | égorgeur | judiciaire |
| armer | empreint | képi |
| arrestation | empreinte | keuf |
| assassin | emprisonnement | kidnapper |
| autopsie | enigme | maison_d'arrêt |
| autopsier | enquête | maison_de_correction |
| autorité | enquêter | malandrin |
| bagne | enquêteur | malfaiteur |
| bandit | envoyer_barreau | médecin_légiste |
| bas_fond_ville | envoyer_verrou | mettre_verrou |
| brigand | épreuve_judiciaire | meurtre |
| C.R.S. | établissement_pénitentiaire | meurtrier |
| cabaler | être_sous_les_verrous | parricide |
| cabaner | être_victime | pénitencier |
| cachot | fait_geste | pénitentiaire |
| canaille | filature | perquisition |
| captivité | flic | perquisitionner |
| capturer | flinguer | piste |
| cause_décès | fonctionnaire | pistolet |
| chenapan | force_public | plainte |
| commissaire | fripon | police |
| commissariat | fripouille | policé |
| commissionnaire | fuite | policeman |
| condamner_à_mort | fusil | policer |
| condé | gangster | policier |
| confrontation | garde_à_vue | pourchasser |
| coupable | garde_champêtre | poursuite |
| course_poursuite | gardien_de_la_paix | preuve |
| courser | gendarmer | prison |
| crapule | gendarmerie | privé |
| crapuleux | gendrarme | revolver |
| crime | homicide | séquestrer |
| criminel | hors-la-loi | sergent |
| déférer | incarcér | soupçon |
| délit | incarcération | soupçonner |
| démasquer | incriminer | stupéfiant |

FIGURE D.7 – Liste des 120 premiers mots et expressions du dictionnaire thématique du Policier

Nous décidons de classifier naïvement et simplement les synopsis suivant cette intensité thématique avec la règle suivante :

$$\textbf{\textit{SI}}\ 12.33 \leq I_{policier}(S)\ \textbf{\textit{ALORS}}\ Genre_{predit}(S) \leftarrow POLICIER.$$

| Déclaré \ Estimé | *NonDrame* | *Drame* |
|---|---|---|
| *NonDrame* | 3997 *(VN)* | 38 *(FP)* |
| *Drame* | 16 *(FN)* | 12 *(VP)* |

FIGURE D.8 – Matrice de confusion sur la prédiction du Policier. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

Nous pouvons voir avec la matrice de confusion D.8 et sur la figure D.9 que cette mesure permet de retrouver 43% des synopsis policier. A partir de ces résultats nous calculons les deux indicateurs que sont la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP + FP} = \frac{12}{12 + 38} = 24\%$$
$$\text{Rappel} = \frac{VP}{VP + FN} = \frac{12}{12 + 16} = 43\%$$

On remarque sur la figure D.9 que les Faux Positifs sont majoritairement dus aux synopsis du genre **Humour**, **Satire**, **Expérimental** et **Drame**.

| TEST_DRAME<br>Genre | FN | FP | VN | VP | TOTAL |
|---|---|---|---|---|---|
| __litterature | 0.0% ( 0) | 0.0% ( 0) | 100% ( 6) | 0.0% ( 0) | 100% ( 6) |
| _argent | 0.0% ( 0) | 0.0% ( 0) | 100% ( 2) | 0.0% ( 0) | 100% ( 2) |
| _educatif | 0.0% ( 0) | 0.3% ( 1) | 99.7% ( 310) | 0.0% ( 0) | 100% ( 311) |
| _provocateur | 0.0% ( 0) | 2.9% ( 3) | 97.1% ( 100) | 0.0% ( 0) | 100% ( 103) |
| Aventure | 0.4% ( 2) | 0.6% ( 3) | 99.0% ( 475) | 0.0% ( 0) | 100% ( 480) |
| Comédie | 0.0% ( 0) | 0.0% ( 0) | 100% ( 12) | 0.0% ( 0) | 100% ( 12) |
| Comédie_musicale | 0.5% ( 1) | 0.0% ( 0) | 99.5% ( 193) | 0.0% ( 0) | 100% ( 194) |
| Conte | 0.0% ( 0) | 0.0% ( 0) | 100% ( 64) | 0.0% ( 0) | 100% ( 64) |
| Documentaire | 0.0% ( 0) | 0.0% ( 0) | 100% ( 130) | 0.0% ( 0) | 100% ( 130) |
| Drame | 0.7% ( 4) | 0.9% ( 5) | 98.2% ( 547) | 0.2% ( 1) | 100% ( 557) |
| Epopée | 0.0% ( 0) | 0.0% ( 0) | 100% ( 15) | 0.0% ( 0) | 100% ( 15) |
| Erotique | 0.0% ( 0) | 2.8% ( 2) | 97.2% ( 69) | 0.0% ( 0) | 100% ( 71) |
| Expérimental | 0.3% ( 3) | 0.8% ( 7) | 98.8% ( 885) | 0.1% ( 1) | 100% ( 896) |
| Fantastique | 0.5% ( 2) | 0.8% ( 3) | 98.7% ( 368) | 0.0% ( 0) | 100% ( 373) |
| Humour | 0.2% ( 3) | 1.2% ( 16) | 98.4% (1266) | 0.2% ( 2) | 100% (1287) |
| Humour_noir | 0.0% ( 0) | 0.0% ( 0) | 100% ( 2) | 0.0% ( 0) | 100% ( 2) |
| Musical | 0.0% ( 0) | 0.8% ( 1) | 99.3% ( 132) | 0.0% ( 0) | 100% ( 133) |
| Policier | 57.1% ( 16) | 0.0% ( 0) | 0.0% ( 0) | 42.9% ( 12) | 100% ( 28) |
| Promotion | 0.0% ( 0) | 0.0% ( 0) | 100% ( 8) | 0.0% ( 0) | 100% ( 8) |
| Propagande | 0.0% ( 0) | 1.6% ( 1) | 98.4% ( 62) | 0.0% ( 0) | 100% ( 63) |
| Publicité | 0.0% ( 0) | 1.4% ( 1) | 98.6% ( 70) | 0.0% ( 0) | 100% ( 71) |
| Satire | 0.3% ( 1) | 2.0% ( 6) | 97.7% ( 300) | 0.0% ( 0) | 100% ( 307) |
| Science-fiction | 0.0% ( 0) | 0.0% ( 0) | 100% ( 22) | 0.0% ( 0) | 100% ( 22) |
| TOTAL | 0.4% ( 32) | 0.9% ( 49) | 98.4% (5038) | 0.3% ( 16) | 100% (5135) |

FIGURE D.9 – Répartition des résultats de la prédiction du Policier suivant le genre déclaré. *Les effectifs sont entre parenthèses. Les effectifs sont supérieurs aux nombres d'observations en raison de réponses multiples (plusieurs genres par synopsis)*

### D.3.3 Thématique de l'Humour

Comme pour la thématique du Drame nous avons créé un dictionnaire thématique de l'Humour qui contient un peu plus de 100 termes dont la figure D.11 représente les 100 premiers termes. Pour vérifier le pouvoir discriminant de la mesure de l'intensité thématique de l'humour nous calculons pour chaque synopsis son intensité thématique puis nous calculons la moyenne des intensités thématiques pour chacune des catégories du genre déclaré sans tenir compte des non-réponses. On voit sur la figure D.10 que cette intensité thématique est peu importante mais significative (les noms des critères discriminants sont encadrés et correspondent à des moyennes significativement différentes de l'ensemble de l'échantillon au risque de 95% (test de student)) dans le cas du genre **Humour**.

| Genre | V7Filtre_DictHumour_I |
|---|---|
| __litterature | 1.78 (0.73 ) |
| _argent | 0.00 (0.00 ) |
| _educatif | 2.43 (0.57 ) |
| _provocateur | 3.74 (1.22 ) |
| Aventure | 1.79 (0.44 ) |
| Comédie | 3.79 (1.48 ) |
| Comédie_musicale | 2.37 (0.46 ) |
| Conte | 2.15 (0.57 ) |
| Documentaire | 1.73 (0.34 ) |
| Drame | 2.75 (0.51 ) |
| Epopée | 0.00 (0.00 ) |
| Erotique | 6.22 (1.56 ) |
| Expérimental | 2.28 (0.49 ) |
| Fantastique | 1.75 (0.36 ) |
| Humour | 3.89 (1.03 ) |
| Humour_noir | 2.95 (2.08 ) |
| Musical | 2.80 (0.61 ) |
| Policier | 0.00 (0.00 ) |
| Promotion | 0.00 (0.00 ) |
| Propagande | 2.80 (0.35 ) |
| Publicité | 0.00 (0.00 ) |
| Satire | 3.35 (0.86 ) |
| Science-fiction | 1.18 (0.25 ) |
| TOTAL | 2.94 (0.66 ) |

FIGURE D.10 – Moyenne des intensités thématiques de l'humour et écart-types entre parenthèses en fonction du genre déclaré

| | | |
|---|---|---|
| absurde | crétin | maladresse |
| absurdité | crétinerie | maladroit |
| amusement | crétinisme | maldonne |
| amuser | cruche | niais |
| amusette | crucherie | niaiserie |
| amusoire | débile | nigauderie |
| andouille | demeuré | pitre |
| âne | dérider | pitrerie |
| ânerie | désennuyer | plaire |
| badin | distraire | plaisanterie |
| baladin | divertir | plaisantin |
| balourdise | drolatique | poisson_d'avril |
| bêta | drôle | polisson |
| bête | drôlerie | polissonnerie |
| bêtiser | égayer | récréatif |
| bêtiser | emmancher | récréer |
| blaguer | fantaisie | réjouir |
| blaguer | farce | rigolade |
| blagueur | farceur | rigolo |
| bouffon | festif | rire |
| bouffonnade | folâtre | risée |
| bouffonnerie | folichon | risible |
| boulette | gaffer | rocambole |
| bourder | gag | saltimbanque |
| boutade | gai | sot |
| bouter | gaieté | sottise |
| burlesque | gaillard | stupide |
| canular | gâteux | stupidité |
| cloche | gausser | sympathique |
| clown | gausserie | taquinerie |
| clownerie | heureux | turlupin |
| comédie | humoriste | vanner |
| comique | humour | vanner |
| con | humouristique | zigoto |
| connerie | idiot | |
| corniaud | idiotie | |
| cornichon | imbécile | |
| couillon | imbécillité | |
| couillonnade | joyeuseté | |
| couillonnerie | ludique | |

FIGURE D.11 – Liste des 100 premiers mots et expressions du dictionnaire thématique de l'Humour

Nous décidons de classifier naïvement et simplement les synopsis suivant cette intensité thématique avec la règle suivante :

$$\textbf{\textit{SI}}\ 3.9 \leq I_{humour}(S)\ \textbf{\textit{ALORS}}\ Genre_{predit}(S) \leftarrow HUMOUR.$$

| Estimé / Déclaré | *NonDrame* | *Drame* |
|---|---|---|
| *NonDrame* | 2661 *(VN)* | 115 *(FP)* |
| *Drame* | 1165 *(FN)* | 122 *(VP)* |

FIGURE D.12 – Matrice de confusion sur la prédiction de l'humour. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

Nous pouvons voir avec la matrice de confusion D.12 et sur la figure D.13 que cette mesure permet de ne retrouver que 9.5% des synopsis dont le genre déclaré est l'humour. A partir de ces résultats nous calculons les deux indicateurs que sont la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP + FP} = \frac{122}{122 + 115} = 52\%$$
$$\text{Rappel} = \frac{VP}{VP + FN} = \frac{122}{12 + 1165} = 9.5\%$$

On remarque sur la figure D.13 que les Faux Négatifs sont nombreux. Cela vient du fait que les films d'animation humoristiques ne portent généralement pas cette marque dans leur synopsis. Cependant lorsque le texte porte les traces d'une histoire humoristique alors le vocabulaire utilisé est spécifique (précision assez bonne de 52%). De plus l'humour est un concept abstrait difficile à saisir qui passe donc par une compréhension du texte.

| TEST_DRAME<br>Genre | FN | FP | VN | VP | TOTAL |
|---|---|---|---|---|---|
| __litterature | 0.0% ( 0) | 16.7% ( 1) | 83.3% ( 5) | 0.0% ( 0) | 100% ( 6) |
| _argent | 0.0% ( 0) | 0.0% ( 0) | 100% ( 2) | 0.0% ( 0) | 100% ( 2) |
| _educatif | 15.4% ( 48) | 5.1% ( 16) | 78.1% ( 243) | 1.3% ( 4) | 100% ( 311) |
| _provocateur | 58.3% ( 60) | 2.9% ( 3) | 29.1% ( 30) | 9.7% ( 10) | 100% ( 103) |
| Aventure | 16.5% ( 79) | 3.5% ( 17) | 77.7% ( 373) | 2.3% ( 11) | 100% ( 480) |
| Comédie | 8.3% ( 1) | 8.3% ( 1) | 75.0% ( 9) | 8.3% ( 1) | 100% ( 12) |
| Comédie_musicale | 5.2% ( 10) | 4.1% ( 8) | 90.2% ( 175) | 0.5% ( 1) | 100% ( 194) |
| Conte | 3.1% ( 2) | 4.7% ( 3) | 90.6% ( 58) | 1.6% ( 1) | 100% ( 64) |
| Documentaire | 6.2% ( 8) | 3.1% ( 4) | 90.0% ( 117) | 0.8% ( 1) | 100% ( 130) |
| Drame | 11.3% ( 63) | 3.4% ( 19) | 84.2% ( 469) | 1.1% ( 6) | 100% ( 557) |
| Epopée | 0.0% ( 0) | 0.0% ( 0) | 100% ( 15) | 0.0% ( 0) | 100% ( 15) |
| Erotique | 19.7% ( 14) | 4.2% ( 3) | 71.8% ( 51) | 4.2% ( 3) | 100% ( 71) |
| Expérimental | 8.7% ( 78) | 3.9% ( 35) | 86.2% ( 772) | 1.2% ( 11) | 100% ( 896) |
| Fantastique | 14.2% ( 53) | 3.2% ( 12) | 81.5% ( 304) | 1.1% ( 4) | 100% ( 373) |
| Humour | 90.5% (1165) | 0.0% ( 0) | 0.0% ( 0) | 9.5% ( 122) | 100% (1287) |
| Humour_noir | 0.0% ( 0) | 0.0% ( 0) | 50.0% ( 1) | 50.0% ( 1) | 100% ( 2) |
| Musical | 36.8% ( 49) | 1.5% ( 2) | 57.9% ( 77) | 3.8% ( 5) | 100% ( 133) |
| Policier | 17.9% ( 5) | 0.0% ( 0) | 82.1% ( 23) | 0.0% ( 0) | 100% ( 28) |
| Promotion | 0.0% ( 0) | 0.0% ( 0) | 100% ( 8) | 0.0% ( 0) | 100% ( 8) |
| Propagande | 30.2% ( 19) | 0.0% ( 0) | 68.3% ( 43) | 1.6% ( 1) | 100% ( 63) |
| Publicité | 4.2% ( 3) | 0.0% ( 0) | 95.8% ( 68) | 0.0% ( 0) | 100% ( 71) |
| Satire | 12.7% ( 39) | 5.2% ( 16) | 80.1% ( 246) | 2.0% ( 6) | 100% ( 307) |
| Science-fiction | 9.1% ( 2) | 0.0% ( 0) | 86.4% ( 19) | 4.6% ( 1) | 100% ( 22) |
| TOTAL | 28.7% (1698) | 2.8% ( 140) | 65.5% (3108) | 3.0% ( 189) | 100% (5135) |

FIGURE D.13 – Répartition des résultats de la prédiction du genre humour suivant le genre déclaré. *Les effectifs sont entre parenthèses. Les effectifs sont supérieurs aux nombres d'observations en raison de réponses multiples (plusieurs genres par synopsis)*

Annexe E

# Annexe chapitre fusion

## E.1 Systèmes flous

### E.1.1 La thématique du Policier

Les informations textuelles (intensité thématique du policier et richesse du synopsis) sont purement numériques elles sont donc transformées en valeurs symboliques par l'utilisation d'ensembles flous. Ainsi le concept de possibilité du genre policier associé à l'intensité thématique du même nom est décrit en utilisant cinq variables linguistiques illustrées par les symboles suivants : *possibilité Très Faible d'être du Policier*, *possibilité Faible d'être du Policier*, *possibilité Moyenne d'être du Policier*, *possibilité Haute d'être du Policier* et *possibilité Très Haute d'être du Policier*. La signification floue de chaque symbole (ou terme linguistique) revient à déterminer le sous-ensemble flou des nombres qu'il représente. Elle est illustrée par sa fonction d'appartenance de type trapézoïdale. La partition floue $F_{Policier}$ de l'univers de discours, $I_{Policier}$, est déterminée par l'ensemble des fonctions d'appartenance aux cinq symboles : $\mu_{TF}$, $\mu_F$, $\mu_M$, $\mu_H$ et $\mu_{TH}$ qui constituent le partitionnement de l'univers de discours $I_{Policier}$ noté $L_{Policier}(I_{Policier})$, et est illustrée par la figure E.1.

La fusion des informations symboliques associées à l'intensité thématique et à la richesse est obtenue par le principe de combinaison/projection utilisant des règles floues. L'ensemble de ces règles peuvent être représentées sous la forme d'une matrice comme sur la figure E.2 où les entrées floues sont représentées en ligne (l'intensité thématique $I_{Policier}$) et en colonne (la richesse) par leurs symboles linguistiques. La variable linguistique de sortie **Policier** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que le synopsis traduise la thématique du Policier. Les valeurs prises par la variable de sortie sont représentées dans chacune des cellules de cette matrice.

Finalement nous décidons de classifier les 5804 synopsis de la base grâce à la règle de classification (voir équation 5.8). Nous comparons les résultats de classification avec le genre déclaré. Si le genre déclaré est le policier alors la classifieur à retrouvé le genre du film sinon il s'est trompé. Nous obtenons la matrice de confusion (voir tableau E.1) où chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe déclarée (ou de référence).

A partir de cette matrice de confusion nous calculons les deux indicateurs que sont la précision et le rappel :

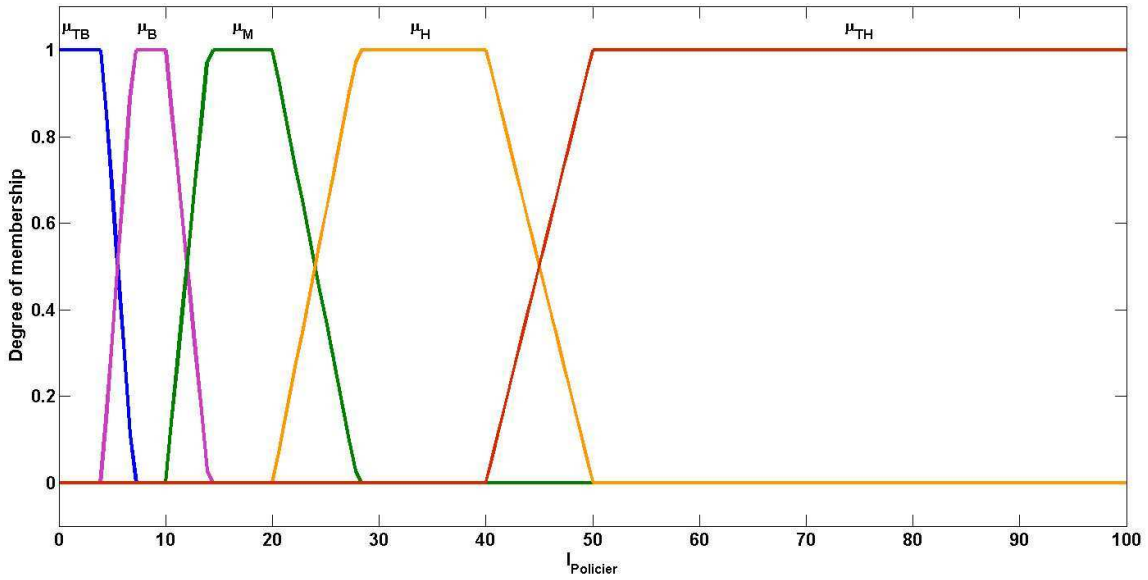



FIGURE E.1 – La partition floue $F_{Policier}$ de l'univers de discours de l'intensité thématique $I_{Policier}$ est déterminée par les fonctions d'appartenance floues : $\mu_{TF}(I_{Policier}) = 1, \forall I_{Policier} \in [0,4]$, $\mu_F(I_{Policier}) = 1, \forall I_{Policier} \in [7,10]$, $\mu_M(I_{Policier}) = 1, \forall I_{Policier} \in [14,20]$, $\mu_H(I_{Policier}) = 1, \forall I_{Policier} \in [28,40]$ et $\mu_{TH}(I_{Policier}) = 1, \forall I_{Policier} \in [50,100]$, (l'axe des ordonnées correspond au degré d'appartenance).

| iPolicier \ Richesse | TC | C | M | L | TL |
|---|---|---|---|---|---|
| TF | *F* | *F* | *F* | *F* | *F* |
| F | *F* | *F* | *F* | *M* | *M* |
| M | *F* | *F* | *H* | *H* | *H* |
| H | *F* | *H* | *H* | *H* | *H* |
| TH | *H* | *H* | *H* | *H* | *H* |

FIGURE E.2 – Règles de combinaison entre l'intensité thématique du Policier et de la richesse du synopsis pour obtenir la mesure du Policier représentée par 3 symboles **F**aible, **M**oyen, **H**aut

$$\text{Précision} = \frac{VP}{VP+FP} = \frac{11}{11+17} = 40\%$$
$$\text{Rappel} = \frac{VP}{VP+FN} = \frac{17}{11+17} = 40\%$$

Nous utilisons également le F-score (ou F-mesure) mesure qui combine la précision et le

| Déclaré \ Estimé | *NonPolicier* | *Policier* |
|---|---|---|
| *NonPolicier* | 4018 *(VN)* | 17 *(FP)* |
| *Policier* | 17 *(FN)* | 11 *(VP)* |

TABLE E.1 – Matrice de confusion sur la prédiction du Policier. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

rappel :

$$F_{score} = 2 * \frac{P * R}{P + R} = 2 * \frac{40 * 40}{40 + 40} = 40\%$$

On remarque que ces taux restent relativement faibles. Cependant l'utilisation et la fusion de l'information de richesse du synopsis a amélioré les résultats de précision par rapport à l'approche basée uniquement sur l'intensité thématique (voir tableau D.8).

### E.1.2 La thématique de l'Humour

Les informations textuelles (intensité thématique de l'humour et richesse du synopsis) sont purement numériques elles sont donc transformées en valeurs symboliques par l'utilisation d'ensembles flous. Ainsi le concept de possibilité du genre humour associé à l'intensité thématique du même nom est décrit en utilisant cinq variables linguistiques illustrées par les symboles suivants : *possibilité Très Faible d'être de l'Humour*, *possibilité Faible d'être de l'Humour*, *possibilité Moyenne d'être de l'Humour*, *possibilité Haute d'être de l'Humour* et *possibilité Très Haute d'être de l'Humour*. La signification floue de chaque symbole (ou terme linguistique) revient à déterminer le sous-ensemble flou des nombres qu'il représente et elle est illustrée par sa fonction d'appartenance de type trapézoïdale. La partition floue $F_{Humour}$ de l'univers de discours, $I_{Humour}$, est déterminée par l'ensemble des fonctions d'appartenance aux cinq symboles : $\mu_{TF}$, $\mu_F$, $\mu_M$, $\mu_H$ et $\mu_{TH}$ qui constituent le partitionnement de l'univers de discours $I_{Humour}$ noté $L_{Humour}(I_{Humour})$, et est illustrée par la figure E.3.

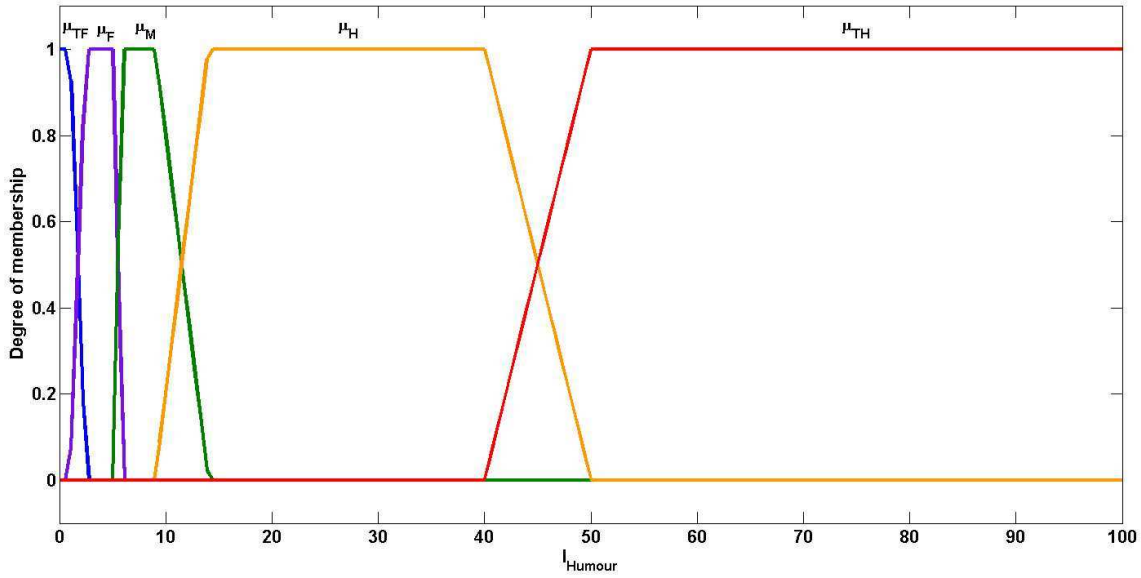



FIGURE E.3 – La partition floue $F_{Humour}$ de l'univers de discours de l'intensité thématique $I_{Humour}$ est déterminée par les fonctions d'appartenance floues : $\mu_{TF}(I_{Humour}) = 1, \forall I_{Humour} \in [0, 1]$, $\mu_F(I_{Humour}) = 1, \forall I_{Humour} \in [2.5, 5]$, $\mu_M(I_{Humour}) = 1, \forall I_{Humour} \in [6, 9]$, $\mu_H(I_{Humour}) = 1, \forall I_{Humour} \in [14, 40]$ et $\mu_{TH}(I_{Humour}) = 1, \forall I_{Humour} \in [50, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

La fusion des informations symboliques associées à l'intensité thématique et à la richesse est obtenue par le principe de combinaison/projection utilisant des règles floues. L'ensemble de ces règles peuvent être représentées sous la forme d'une matrice comme sur la figure E.4 où les entrées floues sont représentées en ligne (l'intensité thématique $I_{Humour}$) et en colonne (la richesse) par leurs symboles linguistiques. La variable linguistique de sortie **Humour** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que le synopsis traduise la thématique de l'Humour. Les valeurs prises par la variable de sortie sont représentées dans chacune des cellules de cette matrice.

Finalement nous décidons de classifier les 5804 synopsis de la base grâce à la règle de

| iHumour \ Richesse | TC | C | M | L | TL |
|---|---|---|---|---|---|
| TF | *F* | *F* | *F* | *F* | *F* |
| F | *F* | *F* | *M* | *H* | *H* |
| M | *F* | *M* | *H* | *H* | *H* |
| H | *M* | *H* | *H* | *H* | *H* |
| TH | *H* | *H* | *H* | *H* | *H* |

FIGURE E.4 – Règles de combinaison entre l'intensité thématique de l'Humour et de la richesse du synopsis pour obtenir la mesure de l'Humour représentée par 3 symboles **F**aible, **M**oyen, **H**aut

classification (voir équation 5.8). Nous comparons les résultats de classification avec le genre déclaré. Si le genre déclaré est le humour alors la classifieur à retrouvé le genre du film sinon il s'est trompé. Nous obtenons la matrice de confusion (voir tableau E.2) où chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe déclarée (ou de référence).

| Déclaré \ Estimé | *NonHumour* | *Humour* |
|---|---|---|
| *NonHumour* | 2652 *(VN)* | 124 *(FP)* |
| *Humour* | 1160 *(FN)* | 127 *(VP)* |

TABLE E.2 – Matrice de confusion sur la prédiction de l'Humour. *Vrai Négatif (VN), Faux Positif (FP), Faux Négatif (FN), Vrai Positif (VP)*

A partir de cette matrice de confusion nous calculons les deux indicateurs que sont la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP+FP} = \frac{127}{127+124} = 51\%$$
$$\text{Rappel} = \frac{VP}{VP+FN} = \frac{127}{127+1160} = 10\%$$

On remarque que ces taux restent relativement faibles et que l'utilisation et la fusion de l'information de richesse du synopsis n'a pas amélioré les résultats de précision ou de rappel (même scores) par rapport à l'approche basée uniquement sur l'intensité thématique (voir tableau D.12).

### E.1.3 Le concept de Froideur

Le concept de Froideur est défini à partir des deux sources d'informations que sont le ratio de couleurs foncées et le ration de couleurs chaudes dans la séquence d'images.

La partition floue $F_{Foncee}$ de l'univers de discours, $R_{Foncee}$, est déterminée par l'ensemble des fonctions d'appartenance aux quatre symboles : $\mu_F$, $\mu_M$, $\mu_H$ et $\mu_{TH}$ qui constituent le partitionnement de l'univers de discours $R_{Foncee}$ noté $L_{Foncee}(R_{Foncee})$, et est illustrée par la figure E.5.

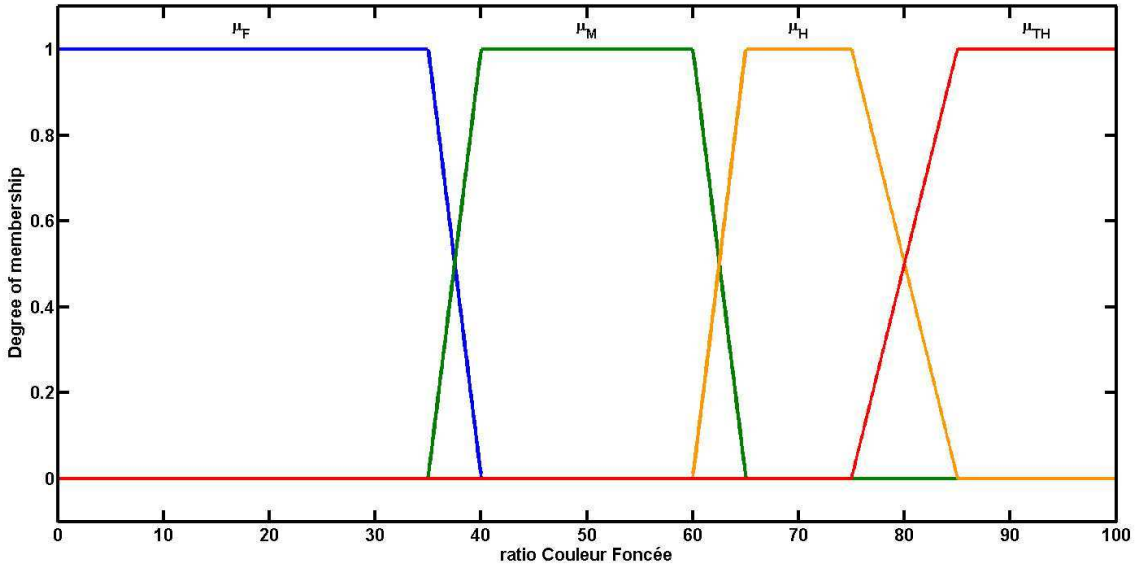



FIGURE E.5 – La partition floue $F_{Foncee}$ de l'univers de discours du ration de couleurs foncées $R_{Foncee}$ est déterminée par les fonctions d'appartenance floues : $\mu_F(R_{Foncee}) = 1, \forall R_{Foncee} \in [0, 35]$, $\mu_M(R_{Foncee}) = 1, \forall R_{Foncee} \in [40, 60]$, $\mu_H(R_{Foncee}) = 1, \forall R_{Foncee} \in [65, 75]$ et $\mu_{TH}(R_{Foncee}) = 1, \forall R_{Foncee} \in [85, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

La partition floue $F_{Chaude}$ de l'univers de discours, $R_{Chaude}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\mu_F$, $\mu_M$ et $\mu_H$ qui constituent le partitionnement de l'univers de discours $R_{Chaude}$ noté $L_{Chaude}(R_{Chaude})$, et est illustrée par la figure E.6.

La fusion de ces informations symboliques est obtenue par le principe de combinaison/projection utilisant des règles floues (voir la figure E.7). La variable linguistique de sortie **Froideur** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que les images traduisent de la froideur. Les valeurs prises par la variable de sortie sont représentées dans chacune des cellules de cette matrice.

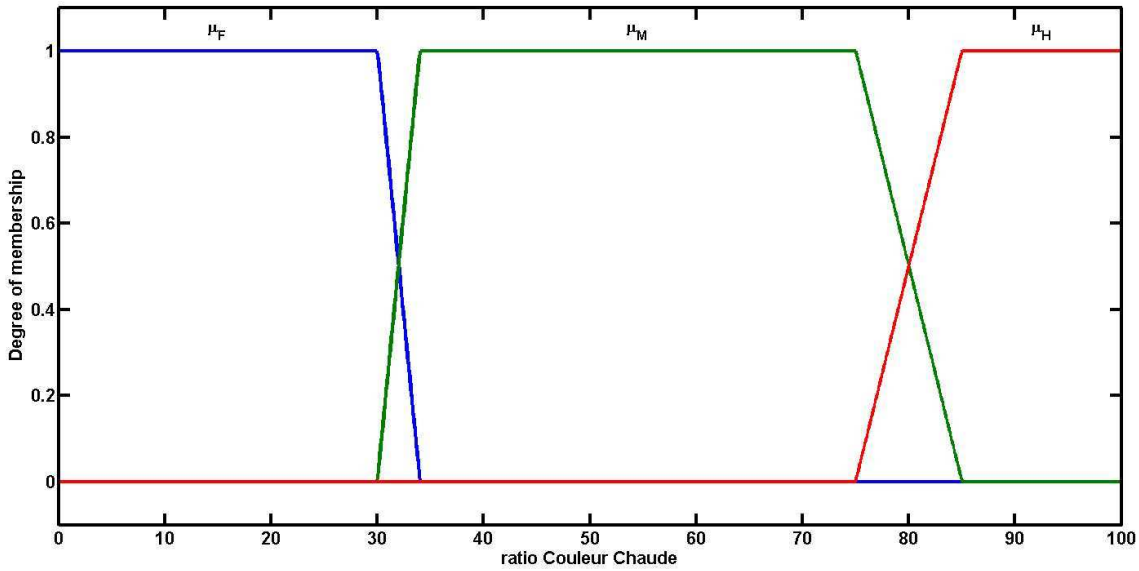



FIGURE E.6 – La partition floue $F_{Chaude}$ de l'univers de discours du ration de couleurs chaudes $R_{Chaude}$ est déterminée par les fonctions d'appartenance floues : $\mu_F(R_{Chaude}) = 1, \forall R_{Chaude} \in [0, 30]$, $\mu_M(R_{Chaude}) = 1, \forall R_{Chaude} \in [34, 75]$ et $\mu_H(R_{Chaude}) = 1, \forall R_{Chaude} \in [85, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

| Couleur Chaude \ Couleur Foncée | F | M | H | TH |
|---|---|---|---|---|
| F | *F* | *F* | *M* | *H* |
| M | *F* | *F* | *F* | *F* |
| H | *F* | *F* | *F* | *F* |

FIGURE E.7 – Règles de combinaison entre le ratio de couleurs foncées et le ration de couleurs chaudes pour obtenir la mesure de la Froideur représentée par 3 symboles **F**aible, **M**oyen, **H**aut

### E.1.4 Le concept de Monotonie

Le concept de film Monotone est défini à partir des deux sources d'informations que sont le ratio de couleurs foncées et la mesure de l'activité globale dans la séquence d'images.

La partition floue $F_{Foncee}$ a été vue juste avant (voir figure E.5).
La partition floue $F_{Activite}$ de l'univers de discours, *Activite*, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\mu_F$, $\mu_M$ et $\mu_H$ qui constituent le partitionnement de l'univers de discours *Activite* noté $L_{Activite}(Activite)$, et est illustrée par la figure E.8.

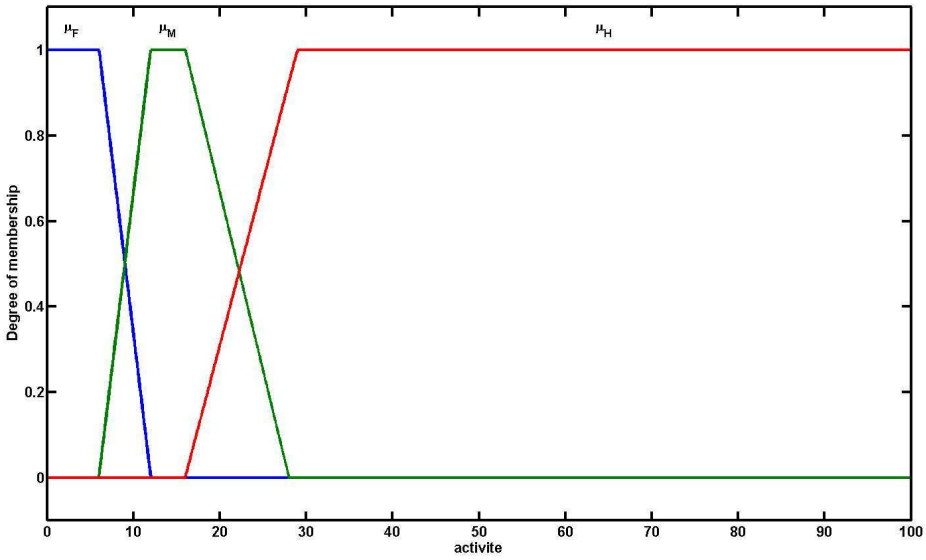



FIGURE E.8 – La partition floue $F_{Activite}$ de l'univers de discours de la mesure d'activité globale *Activite* est déterminée par les fonctions d'appartenance floues : $\mu_F(Activite) = 1, \forall Activite \in [0, 6]$, $\mu_M(Activite) = 1, \forall Activite \in [12, 16]$ et $\mu_H(Activite) = 1, \forall Activite \in [28, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

La fusion de ces informations symboliques est obtenue par le principe de combinaison/projection utilisant des règles floues (voir la figure E.9). La variable linguistique de sortie **Monotone** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que les images traduisent de la monotonie. Les valeurs prises par la variable de sortie sont représentées dans chacune des cellules de cette matrice.

| Activité Globale \ Couleur Foncée | F | M | H | TH |
|---|---|---|---|---|
| F | *F* | *F* | *M* | *H* |
| M | *F* | *F* | *M* | *H* |
| H | *F* | *F* | *F* | *F* |

FIGURE E.9 – Règles de combinaison entre le ratio de couleurs foncées et la mesure d'activité globale pour obtenir le concept de film Monotone représenté par 3 symboles **F**aible, **M**oyen, **H**aut

### E.1.5 Le concept d'Uniformité

Le concept de film Uniforme est défini à partir des deux sources d'informations que sont le ratio de couleurs chaudes et le ratio de variété des couleurs dans la séquence d'images.

La partition floue $F_{Foncee}$ a été vue juste avant (voir figure E.5).
La partition floue $F_{Variation}$ de l'univers de discours, $R_{Variation}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\mu_F$, $\mu_M$ et $\mu_H$ qui constituent le partitionnement de l'univers de discours $R_{Variation}$ noté $L_{Variation}(R_{Variation})$, et est illustrée par la figure E.10.

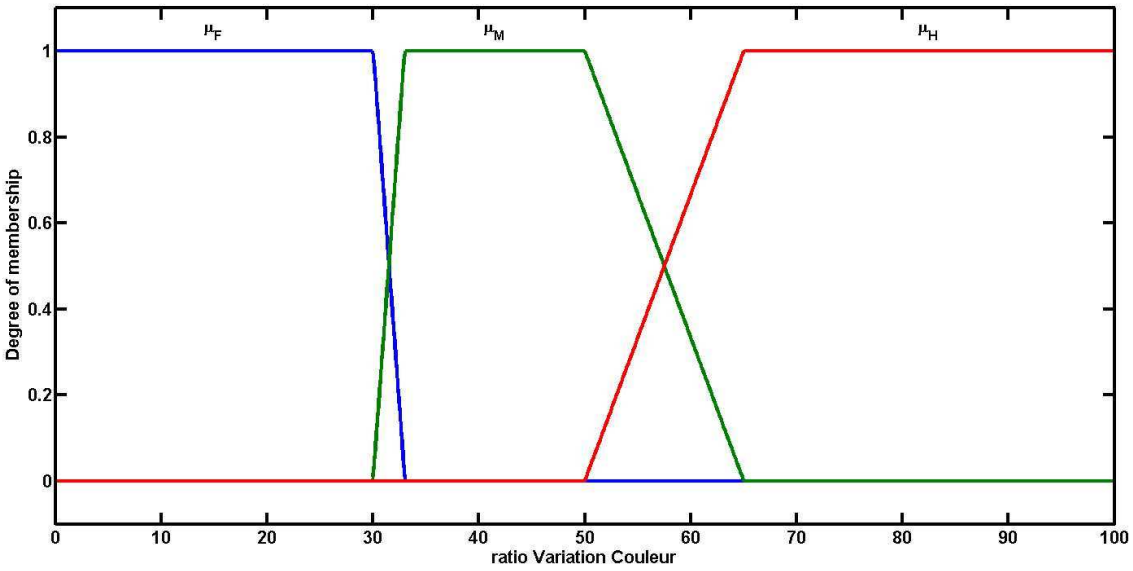



FIGURE E.10 – La partition floue $F_{Variation}$ de l'univers de discours du ration de variation de la palette couleurs $R_{Variation}$ est déterminée par les fonctions d'appartenance floues : $\mu_F(R_{Variation}) = 1, \forall R_{Variation} \in [0, 30]$, $\mu_M(R_{Variation}) = 1, \forall R_{Variation} \in [33, 50]$ et $\mu_H(R_{Variation}) = 1, \forall R_{Variation} \in [65, 100]$, (l'axe des ordonnées correspond au degré d'appartenance).

La fusion de ces informations symboliques est obtenue par le principe de combinaison/projection utilisant des règles floues (voir la figure E.11). La variable linguistique de sortie **Uniforme** est représentée par trois symboles *Faible*, *Moyen*, *Haut* exprimant la possibilité que les images traduisent de l'uniformité. Les valeurs prises par la variable de sortie sont représentées dans chacune des cellules de cette matrice.

| Activité Globale \ Couleur Foncée | F | M | H | TH |
|---|---|---|---|---|
| F | *F* | *F* | *M* | *H* |
| M | *F* | *F* | *M* | *H* |
| H | *F* | *F* | *F* | *F* |

FIGURE E.11 – Règles de combinaison entre le ratio de couleurs foncées et le ratio de variation de la palette couleurs pour obtenir le concept de film Uniforme représentée par 3 symboles **F**aible, **M**oyen, **H**aut

## E.2 La base des 107 films d'animation

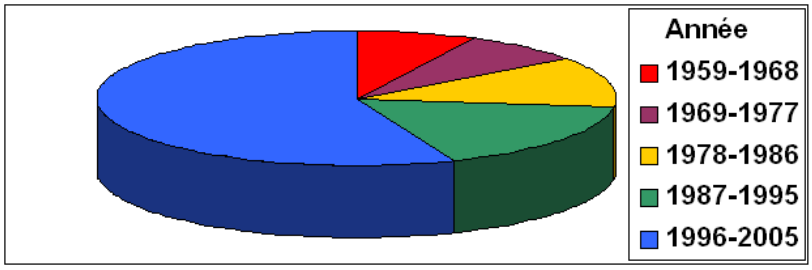



FIGURE E.12 – Répartition des 107 films suivant l'année de production

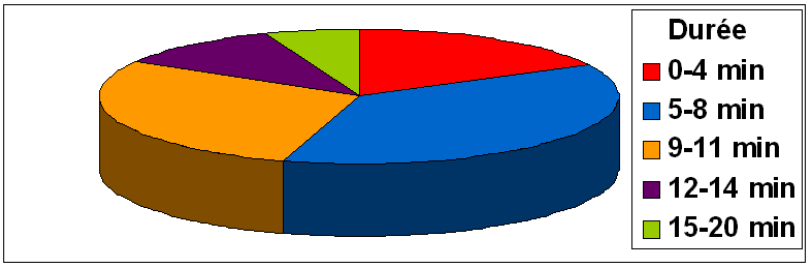



FIGURE E.13 – Répartition des 107 films suivant la durée exprimée en minutes

# Sixième partie

# Bibliographie

# Publications de l'auteur

## Conférences d'audience internationale avec actes

1. G. Païs, F. Deloule, D. Beauchene, P. Lambert, "Animated movie activity characterization by image and text information fusion", 12th International Conference on Information Fusion, CD-ROM , Seattle, USA, July 2009, 8 pages.

2. R. Lagagnière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, B. Ionescu, "Video Summarization from Spatio-Telporal Features", TRECVID BBC Rushes Summarization Workshop, ACM International Conference on Multimedia, Vancouver, Canada, October 2008, pp. 144-148.

## Conférences d'audience nationale et francophone avec actes

3. G. Païs, F. Deloule, D. Beauchene, P. Lambert, "Analyse Texte et Image pour la caractérisation de l'activité dans les Films d'Animation", 27ème congrès INFORSID, CD-ROM , Toulouse, FRANCE, mai 2009, 16 pages.

4. G. Païs, F. Deloule, D. Beauchene, P. Lambert, "Caractérisation de films d'animation pas analyse conjointe texte et image", 9ème journées francophones Extraction et Gestion de Connaissances (EGC-ECOI), Strasbourg, FRANCE, janvier 2009, pp. 43-49.

## Autres conférences

5. D. Beauchene, G. Païs, "Analyse conjointe texte/image pour la caractérisation de films d'animation", Journée scientifique LIMA (Loisirs et Images), Saint-Etienne, FRANCE, juillet 2009.

## Rapports

6. G. Païs, "Analyse Texte et Image pour la caractérisation de Films d'Animation", Rapport interne n° 08/08, LISTIC, 2008.

# Bibliographie


[Abbod *et al.*, 2001] Abbod, M., von Keyserlingk, D., Linkens, D. et Mahfouf, M. (2001). Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems*, 120(2):331–349. *Citée à la page 117.*

[Aigrain *et al.*, 1996] Aigrain, P., Zhang, H. et Petkovic, D. (1996). Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia tools and applications*, 3(3):179–202. *Citée à la page 13.*

[Anderson et Pérez-Carballo, 2001] Anderson, J. et Pérez-Carballo, J. (2001). The nature of indexing : how humans and machines analyze messages and texts for retrieval. Part I : Research, and the nature of human indexing. *Information Processing and Management*, 37(2):231–254. *Citée à la page 5.*

[Arman *et al.*, 1993] Arman, F., Hsu, A. et Chiu, M. (1993). Image processing on compressed data for large video databases. pages 267–272. *Citée à la page 13.*

[Aubin, 2002] Aubin, S. (2002). Grammaire de constituants ou grammaire de dépendance ? quel type d'analyseur choisir pour un système d'extraction d'information ? Mémoire de D.E.A., Institut National de Langues et Civilisations Orientales. *Citée à la page 95.*

[Avilés-Cruz *et al.*, 2005] Avilés-Cruz, C., Rangel-Kuoppa, R., Reyes-Ayala, M., Andrade-Gonzalez, A. et Escarela-Perez, R. (2005). High-order statistical texture analysis : font recognition applied. *Pattern Recogn. Lett.*, 26(2):135–145. *Citée à la page 7.*

[Barra, 2000] Barra, V. (2000). *Fusion d'images 3D du cerveau - Etude de modèles et applications*. Thèse de doctorat, Université d'Auvergne. *Citée à la page 117.*

[Bastière, 1998] Bastière, A. (1998). Methods for multisensor classification of airborne targets integrating evidence theory. *Aerospace Science and Technology*, 2(6):401–411. *Citée à la page 117.*

[Beauchêne et Deloule, 2009] Beauchêne, D. et Deloule, F. (2009). Une expérience de construction d'ontologie. *Journées Francophones sur les Ontologies (JFO), Poitiers, France.* *Citée aux pages 27 and 132.*

[Benzécri et Benzécri, 1980] Benzécri, J. et Benzécri, F. (1980). *Pratique de l'analyse des données*. Dunod Paris. *Citée à la page 74.*

[Berlin et Kay, 1969] Berlin, B. et Kay, P. (1969). *Basic Color Terms : Their Universality and Evolution*. University of California Press. *Citée à la page 51.*

[Bertini *et al.*, 2006] Bertini, M., Del Bimbo, A. et Nunziati, W. (2006). Automatic detection of player's identity in soccer videos using faces and text cues. *In MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, pages 663–666, New York, NY, USA. ACM. *Citée à la page 15.*

[Besada *et al.*, 2004] Besada, J., Molina, J., García, J., Berlanga, A. et Portillo, J. (2004). Aircraft identification integrated into an airport surface surveillance video system. *Machine Vision and Applications*, 15(3):164–171. *Citée à la page 117.*

[Birren, 1969] Birren, F. (1969). Principles of color. *Citée à la page 24.*

[Boggs, 1996] Boggs, J. (1996). *The art of watching films*. Mayfield Publishing Company, 1280 Villa Street, Mountain View, CA 94041. *Citée à la page 12.*

[Bouchon-Meunier, 1993] Bouchon-Meunier, B. (1993). *La logique floue*. Presses universitaires de France. *Citée aux pages 127 and 129.*

[Bouillot, 2008] Bouillot, D. (2008). Pages web de d.bouillot. http ://www.lisiere.com/bouillot.htm. *Citée à la page 179.*

[Bouthemy *et al.*, 1999] Bouthemy, P., Gelgon, M. et Ganansia, F. (1999). A unified approach to shot change detection and camera motion characterization. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1030–1044. *Citée aux pages 9 and 35.*

[Bremond, 1973] Bremond, C. (1973). *Logique du récit*. Éditions du Seuil, Paris. *Citée à la page 90.*

[Brunelli *et al.*, 1999] Brunelli, R., Mich, O. et Modena, C. (1999). A survey on the automatic indexing of video data. *Journal of visual communication and image representation*, 10(2):78–112. *Citée à la page 13.*

[Brunet, 1988] Brunet, É. (1988). Une mesure de la distance intertextuelle : la connexion lexicale. *Revue Informatique et Statistique dans les Sciences humaines (Le nombre et le texte*, pages 1–4. *Citée à la page 73.*

[Brunet, 2000] Brunet, E. (2000). Qui lemmatise dilemme attise. *Lexicometrica*, 2:1–19. *Citée à la page 72.*

[Brunet, 2006] Brunet, E. (2006). Navigation dans les rafales. *Disponible sur : http ://www.cavi. univparis3. fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_ EB. pdf.* *Citée à la page 78.*

[Brunet *et al.*, 2004] Brunet, É., Bases, C. *et al.* (2004). Peut-on mesurer la distance entre deux textes ? *Citée à la page 73.*

[Buehler *et al.*, 2009] Buehler, P., Everingham, M. et Zisserman, A. (2009). Learning sign language by watching TV (using weakly aligned subtitles). *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, volume 2. *Citée à la page 15.*

[Bujor *et al.*, 2002] Bujor, F., Valet, L., Trouve, E., Mauris, G. et Bolon, P. (2002). An interactive fuzzy fusion system applied to change detection inSAR images. *In Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, volume 2. *Citée à la page 117.*

[Bulterman *et al.*, 2007] Bulterman, D. C. A., Jansen, A. J., Cesar, P. et Cruz-Lara, S. (2007). An efficient, streamable text format for multimedia captions and subtitles. *In DocEng '07 : Proceedings of the 2007 ACM symposium on Document engineering*, pages 101–110, New York, NY, USA. ACM. *Citée à la page 15.*

[Caicedo *et al.*, 2008] Caicedo, J., Gonzalez, F. et Romero, E. (2008). Content-based medical image retrieval using low-level visual features and modality identification. *Lecture Notes In Computer Science*, pages 615–622. *Citée à la page 149.*

[Calinski et Harabasz, 1974] Calinski, R. et Harabasz, J. (1974). A dendrite method for cluster analysis. *In Commun. Statistics*, volume 3, pages 1–27. *Citée à la page 62.*

[Carré et Philippe, 2000] Carré, M. et Philippe, P. (2000). Indexation Audio : un état de l'art. *Annals of Telecommunications*, 55(9):507–525. *Citée à la page 13.*

[Chauveau, 2009] Chauveau, j. (2009). La théorie de l'évidence - notes de cours de master 1 informatique université d'angers. http://julien.chauveau.online.fr/m1info/intelligence_artificielle/assets/IA-3-Evidence.pdf. *Citée à la page 125.*

[Choi et Baraniuk, 2001] Choi, H. et Baraniuk, R. (2001). Multiscale image segmentation using wavelet-domain hidden Markovmodels. *IEEE Transactions on Image Processing*, 10(9):1309–1321. *Citée aux pages 7 and 14.*

[Chu *et al.*, 2006] Chu, S., Narayanan, S., Kuo, C. et Mataric, M. (2006). Where am I ? Scene recognition for mobile robots using audio features. *In 2006 IEEE International Conference on Multimedia and Expo*, pages 885–888. *Citée à la page 15.*

[Cinquin et Troccaz, 2009] Cinquin, P. et Troccaz, J. (2009). La chirurgie augmentée à Grenoble. *La Revue pour l'histoire du CNRS, N 24-Automne*. *Citée à la page 117.*

[CITIA, 2009a] CITIA (2009a). Animaquid. http ://www.citia.info/culture-patrimoine-animaquid.html. *Citée à la page 26.*

[CITIA, 2009b] CITIA (2009b). Citia, city of moving images. http ://www.annecy.org. *Citée à la page 21.*

[CITIA, 2009c] CITIA (2009c). Citia recherche. http ://labo.citia.info/. *Citée à la page 22.*

[Corridoni et Del Bimbo, 1995] Corridoni, J. et Del Bimbo, A. (1995). Film semantic analysis. *Proceedings of Computer Architectures for Machine Perception*, pages 202–209. *Citée aux pages 11 and 13.*

[Cotsaces *et al.*, 2006] Cotsaces, C., Nikolaidis, N. et Pitas, I. (2006). Video shot detection and condensed representation. a review. *IEEE signal processing magazine*, 23(2):28–37. *Citée à la page 13.*

[Creutz *et al.*, 2007] Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M. et Stolcke, A. (2007). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Trans. Speech Lang. Process.*, 5(1):1–29. *Citée à la page 18.*

[Cristani *et al.*, 2007] Cristani, M., Bicego, M. et Murino, V. (2007). Audio-visual event recognition in surveillance video sequences. *IEEE transactions on multimedia*, 9(2):257. *Citée à la page 16.*

[Cunningham, 2005] Cunningham, H. (2005). Information extraction, automatic. *Encyclopedia of Language and Linguistics,*, pages 665–677. *Citée aux pages 92 and 93.*

[Curtis *et al.*, 2009] Curtis, J., Baxter, D., Wagner, P., Cabral, J., Schneider, D., Witbrock, M. *et al.* (2009). Methods of Rule Acquisition in the TextLearner System. *In Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read.* *Citée à la page 95.*

[Datta *et al.*, 2008] Datta, R., Joshi, D., Li, J. et Wang, J. Z. (2008). Image retrieval : Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60. *Citée à la page 7.*

[Davenport *et al.*, 1991] Davenport, G., Smith, T. et Pincever, N. (1991). Cinematic principles for multimedia. *IEEE Computer Graphics & Applications*, 11(4):67–74. *Citée aux pages 10 and 42.*

[Del Bimbo, 1999] Del Bimbo, A. (1999). *Visual information retrieval.* Morgan Kaufmann. *Citée à la page 16.*

[Dempster, 1968] Dempster, A. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247. *Citée à la page 124.*

[Desmarais et Moscarola, 2004] Desmarais, C. et Moscarola, J. (2004). Analyse de contenu et analyse lexicale, le cas d'une étude en management public. *In Actes du colloque" L'analyse de données textuelles : De l'enquête aux corpus littéraires.* *Citée à la page 75.*

[Duan *et al.*, 2004] Duan, L.-Y., Xu, M., Tian, Q. et Xu, C.-S. (2004). Nonparametric motion model with applications to camera motion pattern classification. *In MULTIMEDIA '04 : Proceedings of the 12th annual ACM international conference on Multimedia*, pages 328–331, New York, NY, USA. ACM. *Citée aux pages 9 and 35.*

[Dubois et Prade, 1988] Dubois, D. et Prade, H. (1988). *Possibility theory : an approach to computerized processing of uncertainty*. Plenum Press New York. *Citée à la page 126.*

[Dubois et Prade, 2004] Dubois, D. et Prade, H. (2004). On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142(1):143–161. *Citée à la page 118.*

[Ennaji *et al.*, 2003] Ennaji, A., Ribert, A. et Lecourtier, Y. (2003). From data topology to a modular classifier. *International Journal on Document Analysis and Recognition*, 6(1):1–9. *Citée à la page 62.*

[Essid, 2005] Essid, S. (2005). *Classisfication automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. Thèse de doctorat. *Citée aux pages 8 and 13.*

[Estoup, 1916] Estoup, J. (1916). Gammes sténographiques. *Institut Stenographique de France, Paris*. *Citée à la page 71.*

[Faudemay *et al.*, 1998] Faudemay, P., Durand, G., Seyrat, C. et Tondre, N. (1998). Indexing and retrieval of multimedia objects at different levels of granularity. *In Proceedings of SPIE*, volume 3527, page 112. *Citée à la page 5.*

[Fodor et Roubens, 1994] Fodor, J. et Roubens, M. (1994). *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Pub. *Citée à la page 128.*

[Foote, 1999] Foote, J. (1999). An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10. *Citée à la page 8.*

[Foucault et Francois, 2009] Foucault, M. et Francois, A. (2009). General Policy Speech of Prime Ministers and Fiscal Choices in France : "Preach Water and Drink Wine !". *Do They Walk Like They Talk ? : Speech and Action in Policy Processes*, page 131. *Citée à la page 74.*

[Fung, 2003] Fung, G. (2003). The disputed federalist papers : Svm feature selection via concave minimization. *In Proceedings of the 2003 conference on Diversity in computing*, pages 42–46. ACM New York, NY, USA. *Citée à la page 74.*

[Gimel'Farb et Jain, 1996] Gimel'Farb, G. et Jain, A. (1996). On retrieving textured images from an image database. *Pattern Recognition*, 29(9):1461–1483. *Citée à la page 7.*

[Goodman *et al.*, 1997] Goodman, I., Mahler, R. et Nguyen, H. (1997). *Mathematics of data fusion*. Springer. *Citée à la page 116.*

[Grabisch et Perny, 2001] Grabisch, M. et Perny, P. (2001). Agrégation multicritère. *Utilisation de la logique floue, Hermes, Paris*. *Citée aux pages 127 and 128.*

[Greimas, 1966] Greimas, A. (1966). *Sémantique structurale*. Inst. for Litt. *Citée à la page 91.*

[Grishman, 1996] Grishman, R. (1996). Message understanding conference-6 : A brief history. *In Proceedings of the 16th International Conference on Computational Linguistics*, pages 466–471. *Citée à la page 93.*

[Guérin-Pac, 1997] Guérin-Pac, F. (1997). La statistique textuelle. Un outil exploratoire en sciences sociales. *Population*, 52(4):865–887. *Citée à la page 75.*

[Haddad et Chevallet, 2003] Haddad, H. et Chevallet, J.-P. (2003). Utilisation des syntagmes nominaux pour la recherche d'information. Université Jean Moulin LYON. EGC

2003 Journées francophones d'Extraction et de Gestion des Connaissances, Atelier "Fouilles de données et recherche d'informations dans des bases de données multi-média semi-structurées". *Citée à la page 95.*

[Hakenberg *et al.*, 2009] Hakenberg, J., Solt, I., Tikk, D., Tari, L., Rheinländer, A., Ngyuen, Q. L., Gonzalez, G. et Leser, U. (2009). Molecular event extraction from link grammar parse trees. *In BioNLP '09 : Proceedings of the Workshop on BioNLP*, pages 86–94, Morristown, NJ, USA. Association for Computational Linguistics. *Citée à la page 95.*

[Hanna *et al.*, 2009] Hanna, P., Rocher, T. et Robine, M. (2009). A robust retrieval system of polyphonic music based on chord progression similarity. *In SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 768–769, New York, NY, USA. ACM. *Citée à la page 8.*

[Hauptmann *et al.*, 2002] Hauptmann, A. G., Jin, R. et Ng, T. D. (2002). Multi-modal information retrieval from broadcast video using ocr and speech recognition. *In JCDL '02 : Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 160–161, New York, NY, USA. ACM. *Citée à la page 18.*

[Herrera-Boyer *et al.*, 2003] Herrera-Boyer, P., Peeters, G. et Dubnov, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21. *Citée à la page 15.*

[Ionescu, 2007] Ionescu, B. (2007). *Caractérisation Symbolique de Séquences d'images : Application aux Films d'Animation.* Thèse de doctorat, Université de Savoie. *Citée aux pages 24, 37, 38, 40, 41, 42, 58, 143, and 162.*

[Ionescu *et al.*, 2005a] Ionescu, B., Coquin, D., Lambert, P. et Buzuloiu, V. (2005a). Analysis and characterization of animation movies. *ORASIS journées francophones des jeunes chercheurs en vision par ordinateur*, CD-Rom. *Citée aux pages 34 and 38.*

[Ionescu *et al.*, 2005b] Ionescu, B., Coquin, D., Lambert, P. et Buzuloiu, V. (2005b). The influence of the color reduction on cut detection in animation movies. *Actes du 20ème Colloque GRETSI sur le Traitement et l'Analyse du Signal et d'Image*, CD-Rom. *Citée aux pages 38, 52, and 182.*

[Istrate, 2003] Istrate, D. (2003). *Détection et reconnaissance des sons pour la surveillance médicale.* Thèse de doctorat en informatique, Institut National Polytechnique de Grenoble - INPG. *Citée à la page 16.*

[Itten, 1974] Itten, J. (1974). *The art of color : the subjective experience and objective rationale of color.* Wiley. *Citée aux pages 23 and 40.*

[Jeannin et Divakaran, 2001] Jeannin, S. et Divakaran, A. (2001). Mpeg-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):720–724. *Citée à la page 9.*

[Jullien, 2008] Jullien, S. (2008). *Systèmes coopératifs de fusion explicitant les dépendances entre les informations : application à l'interprétation d'images tomographiques 3D et à la sélection de films d'animation.* Thèse de doctorat, Université de Savoie. *Citée aux pages 119, 120, and 122.*

[Jung *et al.*, 2004] Jung, K., In Kim, K. et K. Jain, A. (2004). Text information extraction in images and video : a survey. *Pattern Recognition*, 37(5):977–997. *Citée à la page 18.*

[Kakkonen, 2008] Kakkonen, T. (2008). Robustness evaluation of two ccg, a pcfg and a link grammar parsers. *CoRR*, abs/0801.3817. *Citée à la page 96.*

[Kalampalikis, 2003] Kalampalikis, N. (2003). L'apport de la méthode Alceste dans l'analyse des représentations sociales. *Méthodes d'étude des représentations sociales*, pages 147–163. *Citée à la page 76.*

[Kelly et Judd, 1955] Kelly, K. et Judd, D. (1955). The iscc-nbs color names dictionary and the universal color language (the iscc-nbs method of designating colors and a dictionary for color names). Circular 553, National Bureau of Standards, Washington DC. *Citée à la page 51.*

[Koprinska et Carrato, 2001] Koprinska, I. et Carrato, S. (2001). Temporal video segmentation : A survey. *Signal processing : Image communication*, 16(5):477–500. *Citée à la page 10.*

[Korn, 2006] Korn, B. (2006). Autonomous Sensor-based Landing Systems : Fusion of Vague and Incomplete Information by Application of Fuzzy Clustering Techniques. *In From data and information analysis to knowledge engineering : proceedings of the 29th Annual Conference of the Gesellschaft f "ur Klassifikation eV, University of Magdeburg, March 9-11, 2005*, page 454. Springer. *Citée à la page 117.*

[Kosseim et Lapalme, 1998] Kosseim, L. et Lapalme, G. (1998). Exibum : Un systeme experimental d'extraction d'information bilingue. *Rencontre Internationale sur l'extraction le filtrage et le résumé automatique (RIFRA-98), Sfax, Tunisia*, pages 129–140. *Citée à la page 92.*

[Krzanowski et Lai, 1985] Krzanowski, W. J. et Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *In Biometrics*, volume 44, pages 23–34. International Biometric Society. *Citée à la page 62.*

[Kuncheva et Whitaker, 2003] Kuncheva, L. et Whitaker, C. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207. *Citée à la page 121.*

[Labbé et Labbé, 2001] Labbé, C. et Labbé, D. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Citée aux pages 73 and 74.*

[Labbé, 2002] Labbé, D. (2002). LA LEMMATISATION DES GRANDES BASES DE TEXTES Un exemple : Corneille, Molière et Racine. *L'édition électronique en littérature et dictionnairique, évaluation et bilan.* *Citée à la page 72.*

[Labbé et Monière, 2000] Labbé, D. et Monière, D. (2000). «La connexion intertextuelle. Application au discours gouvernemental québécois». *M. Rajman & J.-C. Chappelier (éds.), JADT*, pages 85–94. *Citée à la page 74.*

[Labbé et Monière, 2008] Labbé, D. et Monière, D. (2008). Je est-il un autre ? *Actes JADT*, 2008:647–656. *Citée à la page 74.*

[Lafon et Muller, 1984] Lafon, P. et Muller, C. (1984). *Dépouillements et statistiques en lexicométrie Travaux de linguistique quantitative.* *Citée à la page 76.*

[Lamalle et Salem, 2002] Lamalle, C. et Salem, A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Actes des JADT 2002*, pages 403–411. *Citée à la page 76.*

[Laptev, 2005] Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123. *Citée aux pages 10 and 36.*

[Larlus, 2008] Larlus, D. (2008). *création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d'objets.* Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG. *Citée à la page 7.*

[Lavee *et al.*, 2009] Lavee, G., Rivlin, E. et Rudzsky, M. (2009). Understanding Video Events : A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions*, 39(5):489–504. *Citée à la page 15.*

[Lazebnik *et al.*, 2006] Lazebnik, S., Schmid, C. et Ponce, J. (2006). Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. *In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. *Citée à la page 7.*

[Lebart et Salem, 1994] Lebart, L. et Salem, A. (1994). Statistique textuelle. *Paris : Dunod,/ c1994*. *Citée aux pages 71, 72, and 74.*

[Lee et Narayanan, 2005] Lee, C. et Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303. *Citée à la page 15.*

[Lefevre *et al.*, 2001] Lefevre, E., Colot, O., Vannoorenberghe, P. et De Brucq, D. (2001). Informations et combinaison : les liaisons conflictuelles. *Revue Traitement du Signal*, 18(3):161–177. *Citée à la page 125.*

[Leonardi *et al.*, 2003] Leonardi, R., Migliorati, P. et Prandini, M. (2003). Semantic indexing of sports program sequences by audio-visual analysis. *In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1. *Citée à la page 10.*

[Li *et al.*, 2002] Li, D., Wong, K., Hu, Y. et Sayeed, A. (2002). Detection, classification and tracking of targets in distributed sensor networks. *IEEE Signal Processing Magazine*, 19(2):17–29. *Citée à la page 117.*

[Lienhart, 2001] Lienhart, R. (2001). Reliable transition detection in videos : A survey and practitiner's guide. *MRL, Intel Corporation, http ://www.lienhart.de/Publications/IJIG_ AUG2001.pdf*. *Citée à la page 12.*

[Lienhart et Wernicke, 2002] Lienhart, R. et Wernicke, A. (2002). Localizing and segmenting text in images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(4):256–268. *Citée à la page 18.*

[Liu *et al.*, 2009] Liu, H., Dai, S., Song, E., Yang, C. et Hung, C.-C. (2009). A new k-view algorithm for texture image classification using rotation-invariant feature. *In SAC '09 : Proceedings of the 2009 ACM symposium on Applied Computing*, pages 914–921, New York, NY, USA. ACM. *Citée aux pages 7 and 34.*

[Liu *et al.*, 2007] Liu, Y., Zhang, D., Lu, G. et Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282. *Citée aux pages 6 and 7.*

[Longrée *et al.*, 2004] Longrée, D., Luong, X. et Mellet, S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé. *Actes des Journées internationales d'Analyse statistique des Données Textuelles*. *Citée à la page 77.*

[Lu, 2001] Lu, G. (2001). Indexing and retrieval of audio : a survey. *Multimedia Tools and Applications*, 15(3):269–290. *Citée à la page 7.*

[Lu *et al.*, 2002] Lu, L., Zhang, H. et Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE transactions on speech and audio processing*, 10(7):504–516. *Citée à la page 15.*

[Lu et Suganthan, 2004] Lu, T. et Suganthan, P. N. (2004). An accumulation algorithm for video shot boundary detection. *Multimedia Tools Appl.*, 22(1):89–106. *Citée aux pages 13, 45, and 179.*

[Madhyastha *et al.*, 2003] Madhyastha, H., Balakrishnan, N. et Ramakrishnan, K. (2003). Event Information Extraction Using Link Grammar. *In 13th International Workshop on Research Issues in Data Engineering : Multi-lingual Information Management (RIDE'03)*. *Citée à la page 95.*

[Manguin *et al.*, 2004] Manguin, J., François, J., Eufe, R., Fesenmeier, L., Ozouf, C. et Sénéchal, M. (2004). Le dictionnaire électronique des synonymes du crisco : un mode d'emploi à trois niveaux. *Les Cahiers du CRISCO*, 17. *Citée à la page 106.*

[Marchand, 2008] Marchand, P. (2008). Analyse lexicométrique d'un genre : la déclaration de politique générale. *proceedings of 9th International Conference on Textual Data statistical Analysis,*, 2:777–785. *Citée à la page 75.*

[Marchand et Monnoyer-Smith, 2000] Marchand, P. et Monnoyer-Smith, L. (2000). Les «discours de politique générale» français : la fin des clivages idéologiques ? *Mots*, 62(1):13–30. *Citée à la page 74.*

[Marszalek *et al.*, 2009] Marszalek, M., Laptev, I. et Schmid, C. (2009). Actions in context. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2929–2936. *Citée à la page 151.*

[Mauris *et al.*, 1996] Mauris, G., Benoit, E. et Foulloy, L. (1996). The aggregation of complementary information via fuzzy sensors. *Measurement*, 17(4):235–249. *Citée aux pages 139 and 140.*

[Maussang, 2005] Maussang, F. (2005). *Traitement d'images et fusion de données pour la détection d'objets enfouis en acoustique sous-marine.* Thèse de doctorat, Université Joseph-Fourier - Grenoble. *Citée à la page 126.*

[Maussang *et al.*, 2008] Maussang, F., Rombaut, M., Chanussot, J., Hétet, A. et Amate, M. (2008). Fusion of Local Statistical Parameters for Buried Underwater Mine Detection in Sonar Imaging. *EURASIP Journal on Advances in Signal Processing*, 2008:19 pages. *Citée à la page 117.*

[Mayaffre, 2007] Mayaffre, D. (2007). L'analyse des données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle. *Lexicométrica*, 7. *Citée à la page 76.*

[McDaniel, 2001] McDaniel, D. (2001). An Information Fusion Framework for Data Integration. *In Proceedings of the 13th Software Technology Conference.* *Citée à la page 121.*

[Menegaz *et al.*, 2007] Menegaz, G., Troter, A. L., Sequeira, J. et Boi, J. M. (2007). A discrete model for color naming. *EURASIP J. Appl. Signal Process.*, 2007(1):113–113. *Citée à la page 51.*

[Miller, 1995] Miller, G. (1995). WordNet : a lexical database for English. *Communications of the ACM*, 38(11):39–41. *Citée à la page 100.*

[Milligan et Cooper, 1985] Milligan, G. et Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179. *Citée à la page 62.*

[Mojsilovic, 2005] Mojsilovic, A. (2005). A computational model for color naming and describing color composition of images. *IEEE Transactions on Image Processing*, 14(5): 690–699. *Citée aux pages 52 and 179.*

[Mokhtarian *et al.*, 1997] Mokhtarian, F., Abbasi, S. et Kittler, J. (1997). Efficient and robust retrieval by shape content through curvature scale space. *Image Databases and Multi-Media Search*, pages 51–58. *Citée aux pages 16 and 34.*

[Molla et Hutchinson, 2003] Molla, D. et Hutchinson, B. (2003). Intrinsic versus extrinsic evaluations of parsing systems. *Citée à la page 96.*

[Mosteller et Wallace, 1984] Mosteller, F. et Wallace, D. (1984). *Applied Bayesian and classical inference : the case of the Federalist papers.* Springer Verlag. *Citée à la page 74.*

[Mourad, 1999] Mourad, G. (1999). La segmentation de textes par l'étude de la ponctuation. *CIDE.99 Conférence Internationale sur le Document Electronique*, pages 155–171. *Citée aux pages 14 and 93.*

[Muller, 1967] Muller, C. (1967). *Étude de statistique lexicale : le vocabulaire du théâtre de Pierre Corneille.* Larousse. *Citée à la page 73.*

[Muslea, 1999] Muslea, I. (1999). Extraction patterns for information extraction tasks : A survey. *In The AAAI-99 Workshop on Machine Learning for Information Extraction.* *Citée à la page 92.*

[Naphide et Huang, 2001] Naphide, H. et Huang, T. (2001). A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151. *Citée à la page 19.*

[Orio, 2006] Orio, N. (2006). *Music retrieval : A tutorial and review.* Now Publishers Inc. *Citée à la page 8.*

[Panagiotakis *et al.*, 2006] Panagiotakis, C., Ramasso, E., Tziritas, G., Rombaut, M. et Pellerin, D. (2006). Shape-motion based athlete tracking for multilevel action recognition. *Lecture Notes in Computer Science*, 4069:385. *Citée à la page 10.*

[Patel et Sethi, 1996] Patel, N. et Sethi, I. (1996). Audio characterization for video indexing. *In Proceedings of SPIE*, volume 2670, page 373. *Citée à la page 13.*

[Patrice *et al.*, 2004] Patrice, C., Lionel, D., Éric, D., Florence, D., Sébastien, S. et Roger, F. (2004). Étude des représentations sociales de la chimiothérapie : une voie d'analyse des relations entre patients et médecins oncologues. *Bulletin du cancer*, 91(3): 279–84. *Citée à la page 75.*

[Perrot, 1980] Perrot, J. (1980). Ponctuation et fonctions linguistiques. *Langue française*, 45(1):67–76. *Citée à la page 12.*

[Petridis et Pantic, 2008] Petridis, S. et Pantic, M. (2008). Fusion of audio and visual cues for laughter detection. *In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 329–338, New York, NY, USA. ACM. *Citée à la page 8.*

[Petrushin, 1999] Petrushin, V. (1999). Emotion in speech : Recognition and application to call centers. *Artificial Neu. Net. In Engr.(ANNIE'99)*, pages 7–10. *Citée à la page 15.*

[Platt, 1999] Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. *Citée à la page 150.*

[Pothos *et al.*, 2007] Pothos, V. K., Theoharatos, C., Economou, G. et Ifantis, A. (2007). Texture retrieval based on a non-parametric measure for multivariate distributions. *In CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 502–509, New York, NY, USA. ACM. *Citée à la page 7.*

[Propp *et al.*, 1970] Propp, V., Derrida, M., Kahn, C., Meletinski, E. et Todorov, T. (1970). *Morphologie du conte.* Seuil. *Citée à la page 90.*

[Pujol, 2009] Pujol, A. (2009). *Contributions à la Classification Sémantique d'Images.* Thèse de doctorat en informatique, Ecole Centrale de Lyon. *Citée aux pages 7 and 16.*

[Pyysalo *et al.*, 2006] Pyysalo, S., Ginter, F., Pahikkala, T., Boberg, J., Järvinen, J. et Salakoski, T. (2006). Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. *International Journal of Medical Informatics*, 75(6):430–442. *Citée aux pages 95 and 96.*

[Quénot, 1996] Quénot, G. (1996). Computation of optical flow using dynamic programming. *In IAPR Workshop on machine vision applications*, pages 249–252. *Citée à la page 10.*

[Ramasso, 2007] Ramasso, E. (2007). *Reconnaissance de séquences d'états par le Modèle des Croyances Transférables. Application à l'analyse de vidéos d'athlétisme.* Thèse de doctorat, Université Joseph-Fourier - Grenoble. *Citée aux pages 36 and 126.*

[Ramesh *et al.*, 2002] Ramesh, G., Maniatty, W. et Zaki, M. (2002). Indexing and data access methods for database mining. *In VIIth ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 02), Madison, Wisconsin, USA. In proceedings*. Citeseer. *Citée à la page 5.*

[Reinert, 1986] Reinert, M. (1986). Un logiciel d'analyse lexicale :(Alceste). *Les Cahiers de l'analyse des données*, 11(4):471–481. *Citée à la page 76.*

[Reinert, 1997] Reinert, M. (1997). «Les Mondes lexicaux des six numéros de la revue "Le Surréalisme au Service de la Révolution"». *Cahiers du centre de recherche sur le surréalisme*, pages 270–302. *Citée à la page 76.*

[Reinert, 2002] Reinert, M. (2002). La tresse du sens et la méthode "Alceste" Application aux Rêveries du promeneur solitaire. *JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles.* *Citée à la page 76.*

[Reinert, 2008] Reinert, M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. *Citée à la page 76.*

[Rivlin et Weiss, 1995] Rivlin, E. et Weiss, I. (1995). Local invariants for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):226–238. *Citée à la page 6.*

[Rombaut, 2001] Rombaut, M. (2001). Fusion : Etat de l'art et perspectives. Rapport technique DSP 99.60.078, IUT de Troyes, laboratoire LM2S-UTT. *Citée aux pages 119 and 123.*

[Ruvolo *et al.*, 2008] Ruvolo, P., Fasel, I. et Movellan, J. (2008). Auditory mood detection for social and educational robots. *In IEEE International Conference on Robotics and Automation, 2008. ICRA 2008*, pages 3551–3556. *Citée à la page 15.*

[Salem, 2004] Salem, A. (2004). Introduction à la résonance textuelle. *7èmes Journées internationales d'Analyse statistique des Données Textuelles, Louvain.* *Citée à la page 76.*

[Salton, 1968] Salton, G. (1968). A comparison between manual and automatic indexing methods. *Citée à la page 5.*

[Sarawagi, 2008] Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, 1(3):261–377. *Citée à la page 92.*

[Satoh *et al.*, 1999] Satoh, S., Nakamura, Y. et Kanade, T. (1999). Name-it : Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35. *Citée à la page 16.*

[Saussure *et al.*, 1922] Saussure, F., Bally, C., Séchehaye, A., Riedlinger, A., Calvet, L. et De Mauro, T. (1922). *Cours de linguistique générale.* Payot, Paris. *Citée à la page 70.*

[Scaringella *et al.*, 2006] Scaringella, N., Zoia, G. et Mlynek, D. (2006). Automatic genre classification of music content : a survey. *IEEE Signal Processing Magazine*, 23(2): 133–141. *Citée à la page 8.*

[Schmid et Mohr, 1997] Schmid, C. et Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5): 530–535. *Citée à la page 6.*

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47. *Citée à la page 74.*

[Shafer, 1976] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press Princeton, NJ. *Citée à la page 124.*

[Sitbon et Bellot, 2005] Sitbon, L. et Bellot, P. (2005). Segmentation thématique par chaînes lexicales pondérées. *TALN 2005*, 1:505–510. *Citée à la page 14.*

[Sleator et Temperley, 1991] Sleator, D. et Temperley, D. (1991). Parsing english with a link grammar. Rapport technique, Carnegie Mellon University Computer Science technical report CMU-CS-91-196. *Citée à la page 95.*

[Smeaton *et al.*, 2009] Smeaton, A., Over, P. et Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid : a 5-Year Retrospective of Achievements. *In Multimedia Content Analysis : Theory and Applications*, pages 151–174. Springer Verlag. *Citée à la page 15.*

[Smeulders *et al.*, 2000] Smeulders, A., Worring, M., Santini, S., Gupta, A. et Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380. *Citée aux pages 6 and 15.*

[Smith *et al.*, 2004] Smith, P., Drummond, T. et Cipolla, R. (2004). Layered motion segmentation and depth ordering by tracking edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):479–492. *Citée à la page 10.*

[Snoek et Worring, 2005] Snoek, C. et Worring, M. (2005). Multimodal video indexing : A review of the state-of-the-art. *Multimedia Tool and Applications*, 25(1):5–35. *Citée aux pages 6, 19, and 20.*

[Snoek *et al.*, 2005] Snoek, C., Worring, M. et Smeulders, A. (2005). Early versus late fusion in semantic video analysis. *In Proceedings of the 13th annual ACM International Conference on Multimedia*, page 402. ACM. *Citée à la page 15.*

[Snoek *et al.*, 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M. et Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. *In MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA. ACM. *Citée aux pages 15 and 16.*

[Sphinx, 2009] Sphinx (2009). Logiciel le sphinx. http ://www.lesphinx-developpement.fr. *Citée aux pages 80 and 106.*

[Sugeno, 1974] Sugeno, M. (1974). *Theory of fuzzy integrals and its applications*. Thèse de doctorat, Tokyo Institute of Technology Tokyo, Japan. *Citée à la page 128.*

[Swets et Weng, 1996] Swets, D. et Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836. *Citée à la page 16.*

[Szummer et Picard, 1998] Szummer, M. et Picard, R. (1998). Indoor-outdoor image classification. *In 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, 1998. Proceedings.*, pages 42–51. *Citée aux pages 14 and 15.*

[Tong et Chang, 2001] Tong, S. et Chang, E. (2001). Support vector machine active learning for image retrieval. *In Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM New York, NY, USA. *Citée à la page 149.*

[Troccaz, 2006] Troccaz, J. (2006). La chirurgie urologique assistée par ordinateur et robot. *Prog. Urol*, 16(2):112–120. *Citée à la page 117.*

[Trohidis *et al.*, 2008] Trohidis, K., Tsoumakas, G., Kalliris, G. et Vlahavas, I. (2008). Multilabel classification of music into emotions. *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*. *Citée aux pages 8 and 163.*

[Trucco et Plakas, 2006] Trucco, E. et Plakas, K. (2006). Video tracking : a concise survey. *IEEE Journal of Oceanic Engineering*, 31(2):520–529. *Citée aux pages 10 and 36.*

[Tsai et Wu, 2008] Tsai, C. et Wu, J. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4):2639–2649. *Citée à la page 149.*

[Tsekeridou et Pitas, 1999] Tsekeridou, S. et Pitas, I. (1999). Audio-visual content analysis for content-based video indexing. *In Proc. of ICMCS*, volume 1, pages 667–672. *Citée à la page 20.*

[Turenne, 2001] Turenne, N. (2001). Etat de l'art de la classification automatique pour l'acquisition de connaissances à partir de textes. Rapport technique, INRA. *Citée à la page 95.*

[Tzanetakis et Cook, 2002] Tzanetakis, G. et Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302. *Citée à la page 15.*

[Ullman *et al.*, 2001] Ullman, S., Sali, E. et Vidal-Naquet, M. (2001). A fragment-based approach to object representation and classification. *Lecture notes in computer science*, pages 85–102. *Citée à la page 7.*

[Vailaya *et al.*, 2001] Vailaya, A., Figueiredo, M., Jain, A., Zhang, H., Technol, A. et Alto, P. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130. *Citée à la page 15.*

[Vailaya *et al.*, 1998] Vailaya, A., Jain, A. et Zhang, H. (1998). On image classification : City images vs. landscapes. *Pattern Recognition*, 31:1921–1936. *Citée à la page 15.*

[Valet, 2001] Valet, L. (2001). *Un système flou de fusion coopérative : application au traitement d'images naturelles*. Thèse de doctorat, Université de Savoie. *Citée aux pages 20, 117, 118, 119, 120, 121, and 123.*

[Valet *et al.*, 2001] Valet, L., Mauris, G. et Bolon, P. (2001). A statistical overview of recent literature in information fusion. *IEEE Aerospace and Electronic Systems Magazine*, 16(3):7–14. *Citée à la page 118.*

[Van de Sande *et al.*, 2008] Van de Sande, K. E., Gevers, T. et Snoek, C. G. (2008). A comparison of color features for visual concept classification. *In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 141–150, New York, NY, USA. ACM. *Citée aux pages 7 and 16.*

[Vapnik, 1996] Vapnik, V. (1996). Structure of statistical learning theory. *Computational Learning and Probabilistic Reasoning*, page 3. *Citée à la page 149.*

[Viprey, 2004] Viprey, J. (2004). «Analyse séquencée de la micro-distribution lexicale». *Actes des Journées internationales d'Analyse statistique des Données Textuelles*. *Citée à la page 78.*

[Visibone, 2009] Visibone (2009). Webmaster palette. http ://www.visibone.com/colorlab/. *Citée à la page 38.*

[Volgyesi *et al.*, 2007] Volgyesi, P., Balogh, G., Nadas, A., Nash, C. et Ledeczi, A. (2007). Shooter localization and weapon classification with soldier-wearable networked sensors. *In Proceedings of the 5th international conference on Mobile systems, applications and services*, page 126. ACM. *Citée à la page 117.*

[Volponi *et al.*, 2003] Volponi, A., Brotherton, T., Luppold, R. et Simon, D. (2003). Development of an information fusion system for engine diagnostics and health management. *In JANNAF 27th airbreathing propulsion subcommittee meeting*. Citeseer. *Citée à la page 117.*

[Vonfelt, 2008] Vonfelt, S. (2008). *La musique des lettres*. Thèse de doctorat, Université de Toulouse-Le Mirail et Université de Parme/Bologne. *Citée à la page 78.*

[Wald, 1999] Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3 Part 1):1190–1193. *Citée à la page 117.*

[Williams *et al.*, 2003] Williams, R., Cotte, O. et Quentin, B. (2003). *Techniques d'animation pour le dessin animé, l'animation 3D et le jeu vidéo*. Eyrolles. *Citée à la page 167.*

[Witten *et al.*, 1999] Witten, I., of Waikato, U. et of Computer Science, D. (1999). *Weka : Practical Machine Learning Tools and Techniques with Java Implementations*. Dept. of Computer Science, University of Waikato. *Citée aux pages 149 and 150.*

[Wold *et al.*, 1996] Wold, E., Blum, T., Keislar, D. et Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE multimedia*, 3(3):27–36. *Citée aux pages 15 and 16.*

[Wu et Zhu, 1999] Wu, Y. et Zhu, J. (1999). A fusion method for estimate of trajectory. *Science in China Series E : Technological Sciences*, 42(2):149–156. *Citée à la page 117.*

[Xiao *et al.*, 2009] Xiao, Z., Dellandréa, E., Dou, W. et Chen, L. (2009). Recognition of emotions in speech by a hierarchical approach. *In International Conference on Affective Computing and Intelligent Interaction (ACII)*. *Citée à la page 8.*

[Xie *et al.*, 2004] Xie, H., Andreae, P., Zhang, M. et Warren, P. (2004). Learning models for english speech recognition. *In ACSC '04 : Proceedings of the 27th Australasian conference on Computer science*, pages 323–329, Darlinghurst, Australia, Australia. Australian Computer Society, Inc. *Citée à la page 18.*

[Xu *et al.*, 2008] Xu, M., Xu, C., Duan, L., Jin, J. S. et Luo, S. (2008). Audio keywords generation for sports video analysis. *ACM Trans. Multimedia Comput. Commun. Appl.*, 4(2):1–23. *Citée à la page 16.*

[Yan et Hauptmann, 2007] Yan, R. et Hauptmann, A. (2007). A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4):445–484. *Citée à la page 18.*

[Zadeh, 1975] Zadeh, L. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information sciences*, 8(3):199–249. *Citée à la page 126.*

[Zentner *et al.*, 2008] Zentner, M., Grandjean, D. et Scherer, K. (2008). Emotions evoked by the sound of music : Characterization, classification, and measurement. *Emotion*, 8(4):494–521. *Citée à la page 15.*

[Zhang et Lu, 2004] Zhang, D. et Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19. *Citée à la page 34.*

[Zhang *et al.*, 1993] Zhang, H., Kankanhalli, A. et Smoliar, S. (1993). Automatic partitioning of full-motion video. *Multimedia Systems*, pages 10–28. *Citée à la page 13.*

[Zhang *et al.*, 1995] Zhang, H., Low, C. et Smoliar, S. (1995). Video parsing and browsing using compressed data. *Multimedia tools and applications*, 1(1):89–111. *Citée à la page 9.*

[Zhang et Kuo, 1999] Zhang, T. et Kuo, C. (1999). Hierarchical classification of audio data for archiving andretrieving. *In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings.*, volume 6. *Citée aux pages 15 and 16.*

[Zhao *et al.*, 2003] Zhao, W., Chellappa, R., Phillips, P. J. et Rosenfeld, A. (2003). Face recognition : A literature survey. *ACM Comput. Surv.*, 35(4):399–458. *Citée aux pages 16 and 34.*

[Zhong *et al.*, 1995] Zhong, Y., Karu, K. et Jain, A. (1995). Locating text in complex color images. *Pattern Recognition*, 28(10):1523–1535. *Citée à la page 18.*

[Zimmermann et Zysno, 1980] Zimmermann, H. et Zysno, P. (1980). Latent connectives in human decision making. *Fuzzy sets and systems*, 4(1):37–51. *Citée à la page 128.*

[Zipf, 1949] Zipf, G. (1949). *Human behavior and the principle of least effort : An introduction to human ecology*. addison-wesley press. *Citée à la page 71.*