Structuration automatique de documents audio

Abdesselam Bouchekif

THÈSE

présentée à l'Université du Maine
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 503
« Sciences et Technologies de l'Information et Mathématiques »

Structuration automatique de documents audio

par

Abdesselam Bouchekif

Soutenue publiquement le 03/11/2016 devant un jury composé de :

Patrice Bellot	Professeur à l'Université de Marseille	Rapporteur
Georges Linarès	Professeur à l'Université d'Avignon	Rapporteur
Pascale Sébillot	Professeur à l'INSA de Rennes	Examinatrice
Camille Guinaudeau	Maître de Conférences à l'Université de Paris Sud	Examinatrice
Yannick Estève	Professeur à l'Université du Maine	Directeur de thèse
Géraldine Damnati	Directrice de recherche, Orange Labs	Co-Encadrante de thèse
Nathalie Camelin	Maître de Conférences à l'Université du Maine	Co-Encadrante de thèse
Delphine Charlet	Directrice de recherche, Orange Labs	Invitée

Orange Labs - Lannion
Laboratoire d'Informatique de l'Université du Maine

Table des matières

Remerciements

Mes premiers remerciements sont très sincèrement adressés à ma co-encadrante, le *Dr. Géraldine Damnati* pour son immense implication tout au long de cette thèse. Je la remercie notamment pour m'avoir soutenu et pour la confiance qu'elle m'a donnée durant la thèse. Pour tout cela, je lui suis extrêmement reconnaissant.

Je remercie également ma co-encadrante le *Dr. Nathalie Camelin* pour son implication dans mon encadrement. Je la remercie également pour avoir eu la « patience » de corriger les premières versions de ce manuscrit. Ses remarques ont toujours été précieuses. Je la remercie, entre autres, d'avoir encadré la thèse à et d'avoir effectué de nombreux déplacements à Lannion pour nos réunions de travail.

Je tiens à remercier chaleureusement mon directeur de thèse le *Pr. Yannick Estève* pour m'avoir donné l'opportunité de découvrir la recherche en m'accueillant dans son laboratoire, le *LIUM*. Malgré ses nombreuses occupations, il a été toujours répondu présent. Qu'il soit ici remercié de son suivi et de son soutien tout au long de ces trois années de thése.

Je remercie *Pr. Patrice Bellot* et *Pr. Georges Linarès* de m'avoir fait l'honneur d'accepter d'être rapporteurs de ma thèse. Je remercie également *Pr. Pascal Sébillot* et *Dr. Camille Guinaudeau* qui m'ont fait l'honneur de présider ce jury lors de ma soutenance.

J'assure également de mon entière reconnaissance ma collègue de bureau *Delphine Charlet*. Merci à elle pour les réflexions et conseils scientifiques ainsi que pour sa gentillesse et sa bonne humeur. Je tiens aussi à remercier les *Pr. Frédéric Béchet* et *Pr. Paul Deléglise* d'avoir accepté de suivre mes travaux de thèse dans le cadre du CST (Comité de Suivi de Thèse).

Mille mercis à *Carole Lailler* pour la relecture enrichissante de ce manuscrit, ainsi que pour son immense participation à la phase d'annotation. Merci à *Kévin Vythelingum* pour les conversations scientifiques intéressantes ainsi pour ses conseils toujours très pertinents.

Je tiens à remercie mes chers amis *Malek Boualem (Orange Labs)*, *Ahmed Bensedik*, *Mohammed El Amine Abderrahim (Université de Tlemcen)* et *Amine Mouhoub (Université Paris Dauphine)*.

Mes remerciements vont également à toutes les personnes que j'ai eu le plaisir de côtoyer pendant ces années parmi lesquelles : *Sahar Ghannay*, *Fethi Bougares*, *Etienne Micoulaut*, *Hakim Amokrane*, *Adrien Bardet*, *Walid Aransa* et *Mercedes Garcia Martinez*.

J'adresse un remerciement à tous mes amis de Lannion (*Yassine Nair Benrekia, Idir Edjakouane, Ghassen Jendoubi, Mohammed Chirk Belhadj, Wafaa Ben Ali, Btissam Errahmadi, Rabeh Guedrez, Asmaa Damoh, Ines Saidi, etc.*) avec qui j'ai partagé les meilleurs moments de ma vie.

Mes dernières pensées iront vers ma famille, et en particulier à mon père, qui m'a permis de poursuivre mes études jusqu'à aujourd'hui. Je lui souhaite un prompt rétablissement. À ma chère maman et à mes frères et sœurs. Mon dernier remerciement est adressé à ma chère épouse *Hadjer*.

Résumé

La structuration thématique est une branche du traitement automatique du langage naturel. Elle permet à l'utilisateur de prendre rapidement connaissance de l'ensemble des thèmes traités dans un document contenant une pluralité de thèmes. La structuration thématique est également utilisée indirectement pour améliorer d'autres applications comme la recherche d'information ou le résumé automatique.

Dans cette thèse, nous décrivons notre système de structuration thématique composé de deux modules complémentaires : celui de *segmentation thématique* et celui de *titrage*. Le premier consiste à détecter les changements de thèmes de telle sorte que la zone entourée par deux frontières (*i.e.* segment) traite d'une seule thématique. L'ensemble des segments retournés est alors identifié par des étiquettes anonymes. Le rôle du seconde module est, quand à lui, d'attribuer un titre à chaque segment thématique.
Les principales contributions de cette thèse portent sur l'enrichissement de la représentation vectorielle de l'émission. Nous proposons deux approches concernant la pondération : l'approche *itérative* et l'approche *structurelle*. La pondération permet de pénaliser les mots non discriminants et de donner plus de poids aux mots importants. Dans les deux approches, les poids sont estimés à partir du contenu lui-même (*intra-document*). Celui-ci est alors considéré comme une collection de documents mono-thème.
Nous introduisons également la notion de la *cohésion de la parole* qui regroupe la distribution des mots et des locuteurs dans le calcul de similarité entre les différentes parties de l'émission. La représentation vectorielle est renforcée par des relations sémantiques ; nous utilisons des relations issues des articles de presse du même jour que l'émission à segmenter. Par ailleurs, nous avons proposé deux nouvelles métriques d'évaluation qui reflètent mieux la qualité de segmentation : *CouvN* et *CouvD*. L'idée est de mesurer la performance du système en terme de nombre de segments correctement retournés plutôt que de s'intéresser au nombre de frontières. Concernant le titrage automatique, l'approche que nous avons proposée consiste à chercher les articles de presse traitant du même thème que celui du segment. Le titre du segment est celui de l'article le plus proche thématiquement. Enfin, nous avons proposé une nouvelle métrique

d'évaluation pour la chaîne complète : de la segmentation au titrage automatique.

Mots-clés : Segmentation thématique, titrage automatique, représentation vectorielle, cohésion de la parole, pondération intra-document, relations sémantiques.

Abstract

Topic structuring is task of Natural Language Processing domain. Thus, the user can get rapidly knowledge of all the topics in a document, which contains some topics. This thesis deals with topic structuring and offers a system based on an architecture. This one is composed of two of complementary elements ; on one hand, the topic segmentation and on the other hand, the titling process.

In this document, we describe our topic structuring system. The topic segmentation detects the topics changes. The goal is to underline the area surrounded by two boundaries (*i.e.* segment), which deals with a single topic. The set of obtained segments is identified by anonymous labels (topic 1, topic 2, ...). The second module have to assign a title to each topic segment. The major contributions of this thesis concern of the enrichment of the vector representation of TV Broadcast News. Two approaches of intra-document weighting are proposed : an *iterative* approach and a *structural* approach. Weighting aims to penalize non-discriminating words and weights more the important words. In both approaches, weights are estimated from the content itself (intra-document), which is considered as a collection of only one topic documents.
We introduce the notion of *speech cohesion* for topic segmentation of a spoken content. The idea is to integrate speakers informations and lexical content to calculate the cohesion value. The vector representation is reinforced by semantic relations ; we use relations coming from articles of newspapers published the same day as the document to segment. Furthermore, a new evaluation metric which reflects better the quality of segmentation is proposed. The goal is to measure system's performances according to the number of segments correctly returned. The second element of our system is automatic titling. The proposed approach consists in searching articles dealing with the same topic as the current segment. The title of the current segment is choosen among all articles : the thematically closest one is selected. Finally, we proposed a new evaluation metric for the whole chain : from the automatic segmentation and automatic titling.

Keywords : Topic segmentation, title assignation, vector representation, speech cohesion, intra-document weighting, semantic relations.

Introduction

Contexte

Récemment, un grand progrès a été réalisé dans le développement d'innovations technologiques du web. Par exemple, il y a 15 ans, les sites d'information n'étaient disponibles qu'à travers les pages *HTML :* du texte, des tableaux, des images. Actuellement, de nombreuses chaînes de télévision proposent gratuitement à leurs téléspectateurs des services de rattrapage *via* leurs sites internet. Cette rediffusion des émissions *via* le web ne cesse d'augmenter le nombre de documents disponibles.

Avec l'explosion du volume de données diffusées sur Internet, il est difficile de trouver l'information la plus pertinente. La navigation peut alors s'avérer longue et fastidieuse, surtout si l'information voulue se trouve dans un document contenant des passages totalement indépendants. En effet, les systèmes de recherche d'information classiques retournent soit des documents pertinents dans leur totalité, soit aucun si l'information cherchée est noyée dans la collection. Il se peut alors que le document retourné ne corresponde pas, dans sa grande majorité, à ce qui a été demandé. Cependant, l'utilisateur cherche à accéder le plus rapidement possible aux parties du document qui l'intéressent le plus. Prenons l'exemple du journal d'information illustré dans la figure 1. Si l'utilisateur s'intéresse à « la déclaration du président Sarkozy sur la candidature de Nathalie Kosciusko-Morizet » ou au « résumé du match Marseille-PSG », le moteur de recherche devra retourner les extraits correspondants R_1 et R_2 de l'émission.

FIGURE 1 – *Recherche d'information par fragment.*

La nature des contenus audiovisuels possède de nombreuses caractéristiques

permettant de les structurer automatiquement de différentes manières. On peut notamment citer les trois suivants :

— ***Structuration en locuteurs :*** elle offre à l'utilisateur la possibilité de connaître l'ensemble des locuteurs présents dans l'émission et d'écouter uniquement les intervenants de son choix.

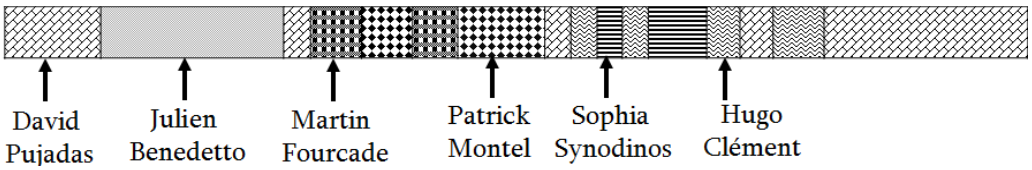


FIGURE 2 – *Structuration en locuteurs.*

La structuration en locuteurs est composée de deux tâches :

- *Segmentation et classification en locuteurs* : cette tâche consiste à déterminer les interventions de chaque locuteur dans un document audio ou vidéo. La segmentation et la classification en locuteurs a attirée beaucoup d'attention et elle reste toujours d'actualité comme le confirme les récentes publications (Lan *et al.*, 2016), (Fredouille et Charlet, 2014) et (Dupuy *et al.*, 2014).
- *l'identification nommée des locuteurs* : cette tâche vise à nommer par prénom et nom les locuteurs apparaissant dans au moins une des modalités portées par la vidéo (Jousse, 2011) et (Béchet *et al.*, 2015).

La figure 3 représente une capture d'écran d'un moteur de recherche des émissions TV basé sur une structuration en locuteurs. Si l'utilisateur saisit

FIGURE 3 – *Moteur de recherche des émissions TV* [1]

la requête : « François Hollande », le système est censé retourner les segments où le syntagme François Hollande a été prononcé et/ou les zones couvrant ses interventions.

— ***Structuration en genres journalistiques*** *:* les informations sont présentées avec différents styles journalistiques (Charlet *et al.*, 2015b) : brève [2], reportage [3], interview [4], *etc.* Cette façon de structurer l'émission permet à l'utilisateur de réduire le temps pour trouver, par exemple, l'interview avec le président de la République ou le débat sur la loi El Khomri.

FIGURE 4 – *Structuration en genres journalistiques.*

— ***Structuration en thèmes*** *:* elle offre à l'utilisateur la possibilité de connaître l'ensemble des thèmes traités durant l'émission. Disposer du début et de la fin de chaque segment thématique permet alors d'ignorer les informations qui ne séduisent pas l'utilisateur. Cette tâche est très prisée dans le traitement automatique du langage naturel. Elle est considérée comme le point de départ de plusieurs applications comme la recherche d'information, le résumé automatique et la modélisation des thèmes.

  Dans cette thèse, nous décomposons la structuration thématique en deux tâches complémentaires : la segmentation thématique (Bouchekif *et al.*, 2015) et le titrage de segments (Bouchekif *et al.*, 2016).

La segmentation thématique consiste à effectuer le pavage d'un document (texte standard, audio ou vidéo) en segments thématiquement homogènes. Ces derniers sont généralement identifiés par des labels anonymes. Le titrage consiste à donner un titre à chaque thème abordé durant l'émission. Cette information est un atout supplémentaire pour une meilleure diffusion des contenus et permet aussi à l'utilisateur de prendre connaissance plus rapidement de l'intégralité des thèmes évoqués.

Nos travaux se sont focalisés sur la structuration en thèmes des documents audio. Nous traitons plus particulièrement les journaux télévisés. Les deux tâches (segmentation thématique et titrage) sont claires dans leur globalité. Cependant, elles nécessitent de définir la notion centrale du problème « *qu'est-ce qu'un*

2. C'est une information courte qui résume un fait. Elle doit répondre aux cinq questions : *où, qui, quoi, quand et pourquoi.*
3. Retrace en images et en son un événement.
4. Entretien d'un journaliste avec une personne (questions-réponses)

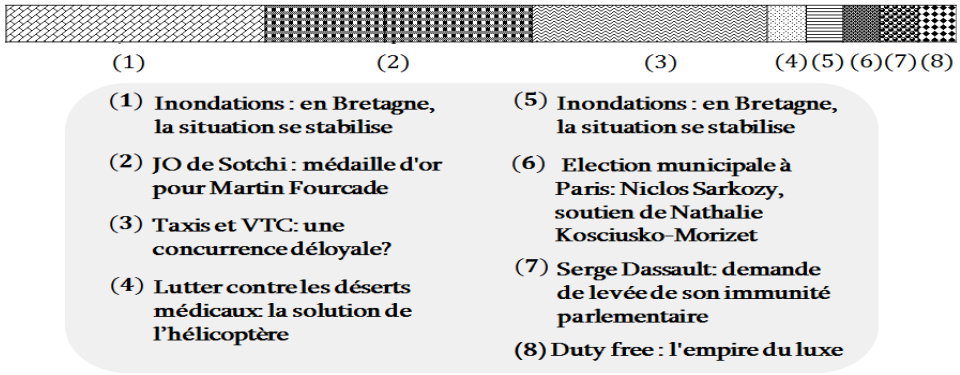


FIGURE 5 – *Structuration en thèmes.*

thème ? ».
Pour ce qui est des journaux télévisés, la détermination des segments thématiques dépendra des consignes données aux annotateurs. Il est donc recommandé de bien établir les règles à suivre surtout pour les cas ambigus (des segments traitant de thématiques proches). C'est le cas par exemple de ces deux thèmes consécutifs « *Ce que pensent les Grecs de l'arrivée au pouvoir de Syriza* » et « *Qui est Alexis Tsipras, le nouveau premier ministre grec ?* » qui peuvent être considérés comme un seul thème ou deux différents. Notre définition du thème est donnée dans la section 1.2.

Problématique et positionnement

La structuration automatique des contenus multimédia en segments homogènes est l'un des problèmes centraux du domaine du traitement automatique de la langue parlée. Les attentes liées à la thèse concernent la mise en œuvre d'un système qui segmente automatiquement les journaux télévisés selon les thèmes abordés et attribue un titre à chaque segment thématique retourné. Cette façon de structurer l'émission peut aider l'utilisateur à accéder plus rapidement aux parties de l'émission qui l'intéressent. Ainsi, elle est considérée comme une brique essentielle de plusieurs systèmes comme les systèmes de traduction automatique et les moteurs de recherche.

Étant donné que les journaux télévisés sont des documents multimodaux, la combinaison des indices vidéo et audio permet d'avoir un système beaucoup plus performant (Wang *et al.*, 2012), (Dumont et Quénot, 2012), *etc.* Cependant, les indicateurs provenant de la vidéo ne sont pas disponibles dans tous les journaux télévisés. Par conséquent, le système perd la capacité à traiter efficacement n'importe quel journal.

L'un des objectifs de ce travail est que le système final soit générique, c'est-à-dire qu'il ne dépende pas de règles éditoriales propres à chaque chaîne. Pour cela, nous avons fait le choix de privilégier certains indices de l'audio et d'écarter les informations visuelles. Nous avons construit un corpus diversifié de journaux télévisés : nos émissions sont issues de 10 chaînes françaises. Ces dernières essaient toujours de mettre en place des règles éditoriales innovantes tout en facilitant la transmission de l'information. Nous souhaitons mettre en place un système de segmentation capable de s'adapter à la diversité des formats. Il nous apparaît que les systèmes de segmentation à l'état de l'art n'ont pas été évalués sur des émissions de différents formats. Il est probable que cela soit dû à la non disponibilité de corpus gratuits.

L'état de l'art propose un nombre important d'algorithmes de segmentation basés sur le calcul de la cohésion lexicale, les plus connus sont : *TextTiling* (Hearst, 1997), *C99*(Choi, 2000) et *MinCut*(Malioutov et Barzilay, 2006). Ces algorithmes se basent essentiellement sur la répétition des mots pour mesurer la similarité entre les différentes parties de l'émission. Même si le degré de sophistication et d'efficacité varie d'un algorithme à un autre, la faible répétition des mots peut perturber les systèmes de segmentation thématique. Pour cette raison, nous nous focalisons davantage sur la représentation vectorielle afin d'assurer la répétition des termes. Nos travaux sont basés sur l'algorithme *TextTiling* qui est largement utilisé comme point de départ pour d'autres systèmes de segmentation thématique.

En ce qui concerne le titrage, l'objectif fixé est non seulement de proposer pour chaque segment un titre correct reflétant l'essentiel de l'information mais aussi de piquer la curiosité de l'utilisateur. Malheureusement, à l'heure actuelle, les techniques liées à cette tâche (comme la compréhension automatique de la langue) ne permettent pas de mener à bien le titrage automatique par génération d'un titre inédit. Afin de pallier ce problème, nous proposons d'exploiter des titres provenant d'articles de presse, écrits par des journalistes professionnels.

Organisation du manuscrit

Ce mémoire est organisé en trois parties. La *partie I* est composée de trois chapitres. Dans le premier, nous reprenons des notions de base liées à la segmentation thématique et le titrage et nous présentons les méthodes de segmentation les plus citées. Le chapitre 2 décrit les corpus que nous avons utilisés pour tester et valider nos contributions. Ensuite, nous présentons les métriques d'évaluation de l'état de l'art ainsi que celles que nous proposons. Le chapitre

3 est consacré à la présentation de notre système de base. Nous donnons également les premiers résultats obtenus.

Les chapitres 4, 5 et 6 forment la *partie II* du manuscrit. Dans le chapitre 4, nous proposons deux approches de pondération *intra-document*. Les chapitres 5 et 6 visent à enrichir la représentation vectorielle du document. Dans un premier temps, nous proposons de considérer conjointement la distribution de mots et de locuteurs. Dans un second temps, nous proposons d'intégrer des relations sémantiques entre les mots dans le processus de la segmentation thématique.

La troisième partie contient un seul chapitre. Ce dernier est dédié à la tâche du titrage automatique qui vise à attribuer un titre à chaque segment.
Quelques éléments de conclusion ainsi qu'une mise en perspective de nos travaux termineront ce document.

Première partie

Segmentation thématique de documents audio

Chapitre 1

État de l'art

Sommaire

1.1 Introduction

Avec les dernières innovations technologiques, les sites des chaînes TV proposent gratuitement à leurs téléspectateurs des services de TV de rattrapage (*Replay/catch up TV*) *via* Internet. Ces services donnent à l'utilisateur la possibilité de voir les émissions des chaînes TV à travers les podcasts. Les vidéos des journaux d'information télévisés intéressent non seulement les utilisateurs mais aussi les fournisseurs d'actualités comme *Google Actualités*, *Yahoo Actualités*, *Orange Actualités*, *etc.* Ils collectent des informations de différentes sources (sites web des journaux, radios, chaînes TV) en agrégeant *automatiquement* les documents de thèmes similaires. La figure 1.1 illustre ce propos. Les articles de presse et les segments vidéo portant sur « carambolage sur l'A13 » ont été

regroupés. Ce regroupement permet à l'utilisateur de visualiser une même information présentée *via* différents médias, ce qui favorise l'accès au pluralisme d'information.

FIGURE *1.1* – *Information traitée par plusieurs sources web, la capture d'écran a été prise à partir de la page d'accueil de Google Actualités.*

Pour permettre ce regroupement, les documents (en particulier les contenus audiovisuels multi-thèmes) doivent être préalablement segmentés en fragments thématiquement homogènes *i.e.* traitant d'un seul thème. Cette tâche est nommée *segmentation thématique*.
En tant qu'application directe, la segmentation thématique permet à l'utilisateur d'accéder plus rapidement aux parties de l'émission qui l'intéressent le plus. Par exemple, la chaîne France 2 enrichit les vidéos du journal télévisé par des marqueurs indiquant les changements de thèmes (voir figure1.2). Cette annotation est effectuée manuellement.
L'intérêt de la segmentation thématique ne se restreint pas à des fins de navigation. En effet, cette tâche a été utilisée dans de nombreux domaines comme la recherche d'information (Prince et Labadié, 2007), le résumé automatique (Silber et McCoy, 2000) ou encore le regroupement de segments thématiquement homogènes (Guinaudeau, 2011).

— En *recherche d'information*, la phase d'indexation est primordiale, elle est basée sur la fréquence des mots. Cependant, la plupart des moteurs de recherche ne prennent pas en compte l'existence des documents multi-thèmes, c'est-à-dire traitant plusieurs thèmes comme les livres, les journaux d'information, *etc.* une grande partie ne correspond pas aux besoins de l'utilisateur. Pour cette raison, il est recommandé au système de répondre très précisément à la requête en évitant de présenter à l'utilisateur le document en entier.
Dans le même contexte, la segmentation thématique est aussi exploitée par les instituts chargés d'archiver les productions audiovisuelles. Dans (Guinaudeau, 2011), l'auteur indique qu'une partie des émissions archivées par l'Institut National de l'Audiovisuel (INA) a été indexée par des

FIGURE 1.2 – *Exemple d'un JT enrichi de la chaîne France2, la capture d'écran a été prise à partir du site de francetvinfo.fr*

documentalistes (*i.e* que l'indexation a été faite manuellement) en produisant un résumé et une indexation thématique de l'émission. Avec le nombre très important de documents multimédias produits chaque année, seules les méthodes automatiques permettent d'exploiter l'ensemble des données archivées.

— En *résumé automatique*, la segmentation thématique est considérée comme une première étape. Par la suite, le système de résumé automatique peut s'appliquer sur les segments et non pas sur le document entier.

— *Regroupement des segments thématiquement homogènes :* comme il a été mentionné précédemment, la segmentation thématique peut être utilisée pour agréger les documents de thèmes similaires. En ce qui concerne les journaux d'information télévisés, cette tâche permet à l'utilisateur d'obtenir un même thème présenté par différentes chaînes. Un tel regroupement est aussi bénéfique sur d'autres types de documents personnels comme les e-mails.

Les premiers systèmes de segmentation thématique se sont focalisés sur la distribution de mots présents dans le document à segmenter. Un changement important de vocabulaire est considéré comme un indice de changement de thème. Cette catégorie de méthodes a été essentiellement dédiée aux documents écrits. La segmentation thématique des documents audiovisuels a pleinement profitée de l'évolution remarquable des techniques d'extraction d'informations à partir de l'audio et de la vidéo (segmentation en locuteur, détection et reconnaissance des textes incrustés, repérage du présentateur principal, *etc.*). Ceci a donné naissance à de nouveaux systèmes regroupant différentes sources d'in-

formation (multimodalité).

Dans ce chapitre, nous commençons d'abord par donner notre définition d'un thème. Ensuite, nous introduisons quelques notions générales sur la segmentation thématique, en abordant plus particulièrement les pré-traitements, la représentation vectorielle d'un document et les mesures de similarité. Puis, nous décrivons l'état de l'art des méthodes de segmentation thématique les plus utilisées. Enfin, nous présentons quelques corpus qui ont été utilisés pour évaluer les systèmes de segmentation thématique.

1.2 Qu'est ce qu'un thème ?

Le thème est la notion centrale du travail présenté dans cette thèse. Les communautés du traitement de la langue écrite et parlée n'ont pas trouvé un consensus car la définition d'un thème est essentiellement liée au domaine d'application et aux résultats attendus. Néanmoins, nous pouvons considérer que la définition d'un thème est liée au contenu sémantique d'un document.

La modélisation thématique (en anglais *topic modeling*) permet d'extraire les thèmes saillants d'une collection de documents textuels. Dans ce domaine, la notion d'un thème est fortement liée au nombre de thèmes attendu (Gaussier et Yvon, 2011). Il est possible de définir différents niveaux de granularité. Par exemple, nous pouvons mettre en évidence trois niveaux de granularité : *générique*, *spécifique* et *fine*. Ainsi, si le but est de classer les documents selon un premier niveau de domaines représentant chacun une très large couverture de documents, les thèmes seront considérés comme génériques. Par exemple, *Google Actualités* regroupe les articles de presse selon des grands domaines prédéfinis manuellement : *International, France, Entreprises, Science/Tech, Sports, Culture et Santé*. Les thèmes deviennent spécifiques lorsque l'on souhaite répartir les documents appartenant à un même thème générique en fonction de leur proximité sémantique. Ainsi, *Google Actualités* proposera le thème spécifique *Élections législatives grecques de 2015* pour le thème générique *International* et le thème spécifique *Inondations en Bretagne* pour le thème *France*.
Le niveau de granularité d'un thème peut encore s'affiner : le thème spécifique *Inondations en Bretagne* peut contenir *(1) Intempéries : cinq départements en alerte orange inondation, (2) Inondations à Quimperlé. Habitants inondés, secours et élus excédés , (3) Inondations. Geilenkirchen offre 9.000EUR, etc.*

Dans les travaux autour de la segmentation thématique, (Labadié, 2008) définit un thème comme étant l'information centrale sur laquelle s'articule un acte de communication.
La structuration thématique des émissions télévisées nécessite une définition

claire d'un thème. En effet, sans un préalable accord entre les annotateurs, il est possible d'avoir plusieurs segmentations de référence pour une même émission. (Guinaudeau, 2011) considère un reportage (éventuellement associé à ses plateaux de lancement et de fin) comme étant un segment thématique. Si cette définition peut s'appliquer pour des journaux télévisés classiques, la fabrication d'un journal varie d'une chaîne à l'autre et la notion de reportage n'est alors pas toujours suffisante.

Certaines chaînes, comme *TF1* ou *France 2*, diffusent les journaux télévisés sur leurs sites web en offrant à l'utilisateur la possibilité de consulter leur contenu partiellement (*i.e* par thème) ou intégralement. L'exemple de la table 1.1 présente une segmentation thématique effectuée *manuellement* par la chaîne France 2 [1]. On constate que le niveau de granularité utilisé est très fin. Par exemple, l'évènement des élections en Grèce est présenté en 6 segments thématiques dont chacun renvoie une information particulière.

Thème 1	Victoire de Syriza : l'espoir renaît en Grèce.
Thème 2	Grèce : Des négociations pour former un gouvernement de coalition.
Thème 3	Syriza : utopie économique ou programme réaliste ?
Thème 4	Grèce : pour appliquer son programme, Alexis Tsipras va devoir renégocier la dette.
Thème 5	Le FN et la gauche saluent la victoire de Syriza, la droite divisée.
Thème 6	La dette grecque en chiffres.
Thème 7	De nouvelles mesures pour lutter contre l'alcool au volant.
Thème 8	Sécurité routière : vers une réduction de la vitesse en ville ?
Thème 9	Ouverture du procès de l'affaire Bettencourt.
Thème 10	Loi Macron : les points qui fâchent.
Thème 11	L'ouverture des magasins le dimanche ne fait pas l'unanimité dans les communes.
Thème 12	New York se prépare à affronter l'une des plus violentes tempêtes de neige de son histoire.
Thème 13	Des lycéens visitent le camp d'extermination d'Auschwitz-Birkenau.
Thème 14	Demis Roussos, légendaire voix grecque de la variété internationale.
Thème 15	A la découverte de l'univers de Cézanne, à Aix-en-Provence.
Thème 16	Sarah et Déborah Nemtanu : deux soeurs virtuoses du violon réunies sur un même album.

***Tableau 1.1** – L'ensemble des Thèmes traités durant le journal 13Heures de France2 diffusé le 26/01/2015.*

Dans cette thèse, nous avons adopté la définition du thème avec la granularité

1. http ://www.francetvinfo.fr/replay-jt/france-2/13-heures/jt-de-13h-du-lundi-26-janvier-2015_801057.html

la plus fine comme celle pratiquée manuellement par les chaînes de télévision. Un thème est alors défini comme une unité :

— Focalisée sur une information précise, qui se déroule à un instant et un endroit donnés.

— Utilisable dans d'autres tâches automatiques comme la recherche d'information, le résumé, le titrage automatique, *etc.* Par exemple, si un utilisateur s'intéresse uniquement à la réaction de la gauche et du FN vis-à-vis des élections grecques, le moteur de recherche doit retourner le *thème 5* de l'émission (voir le tableau 1.1).

Pour bien illustrer la notion de thème, nous présentons dans le tableau 1.2, deux extraits du journal de 13 Heures de France 2 du *13/02/2014*. Le fait que les trois

Extrait *n°1*	Redon : l'inquiétude face aux précipitations.
	Grande-Bretagne : 2000 soldats déployés contre les inondations.
	Etats-Unis : nouvelle tempête de neige.
...	...
Extrait *n°2*	JO de Sochi : déception côté français.
	Sochi : un climat trop chaud ?

Tableau 1.2 – *Deux extraits du journal 13 Heures du 13/02/2014.*

premiers segments thématiques portent sur des événements ayant lieu dans des endroits différents (France, Grande-Bretagne et Etat-Unis) impliquent qu'ils ne peuvent pas être regroupés au sein d'un même thème. De plus, chaque thème traite une information particulière.
Même si les segments thématiques du deuxième extrait portent tous les deux sur les JO de *Sochi*, ils sont malgré tout classés en deux thèmes différents. En effet, le premier informe sur la participation de la France, tandis que le deuxième parle du climat.

1.3 Notions générales

Avant de présenter les méthodes de segmentation thématique à l'état de l'art, nous donnons quelques notions qui seront utiles par la suite, en abordant particulièrement les pré-traitements, la représentation vectorielle et la mesure de similarité entre deux vecteurs. Ces notions sont utilisées systématiquement et indépendamment de la méthode utilisée.

1.3.1 Pré-traitements

Par nature, les données brutes contiennent plusieurs sources de bruit. L'étape de pré-traitement de données textuelles consiste à les préparer pour un traitement automatique efficace. Le pré-traitement de données est considéré comme une opération préliminaire et primordiale dans plusieurs domaines comme la recherche d'information et la segmentation thématique. Nous rappelons ici deux pré-traitements classiques :

Lemmatisation : les mots d'une langue donnée sont accordés en genre, en nombre et en mode (indicatif, impératif...). Le rôle d'un lemmatiseur est de ramener le mot à sa forme canonique (*i.e* les verbes à l'infinitif et les autres mots au masculin singulier). Par exemple, la forme canonique des mots *petit, petite, petits et petites* est *petit*. Plusieurs outils de lemmatisation sont disponibles : *Lia-tagg* [2], *TreeTagger* (Schmid, 1994), *Macaon* (Nasr *et al.*, 2011), *etc.* Ce processus permet de réduire la taille du vocabulaire, de faire apparaître la répétition cachée des mots et par conséquent améliore la qualité du système de segmentation thématique.

Filtrage des mots outils : dans le langage naturel, pour qu'une phrase soit compréhensible, elle doit contenir des noms, adjectifs, verbes, adverbes et des mots fonctionnels. Ces derniers sont des mots non porteurs de sens par rapport à leurs catégories grammaticales. En français, les mots : *le, la, de, du, ce, etc.* et les auxiliaires *être* et *avoir* sont ainsi présents dans presque tous les documents, quel que soit le thème considéré (Amini et Gaussier, 2013). Dans certains domaines comme la recherche d'information et la segmentation thématique, ces mots sont considérés comme une source de bruit et leur suppression est primordiale. La plupart des algorithmes de segmentation utilisent des *stop-listes* (Malioutov et Barzilay, 2006), (Choi, 2000) adaptées selon la nature des données.

1.3.2 Représentation vectorielle d'un document

La représentation vectorielle est une phase primordial pour manipuler les documents textuels. Elle consiste à représenter un document sous la forme d'un vecteur de mots. Le processus le plus simple est de compter le nombre d'occurrences des mots dans les documents et de les reporter dans un vecteur de mots. En segmentation thématique, la méthode la plus répondue consiste à transformer le document en une matrice de taille $m \times n$ où m est le nombre d'unités de base [3] considérées et n est la taille du vocabulaire. L'élément $e_{i,j}$ de la matrice représente le nombre d'occurrences ou le poids du $i^{ème}$ terme dans la $j^{ème}$ *unité*.

2. http ://lia.univ-avignon.fr/chercheurs/bechet/download_fred.html
3. La définition de l'unité de base est donnée ci-dessous.

Choix de l'unité de base
Les systèmes conçus pour segmenter un texte standard ((Hearst, 1997), (Choi, 2000), *etc.*) considèrent soit un paragraphe ou une phrase comme *unité de base*. La notion de phrase ne correspond à rien lorsque l'on travaille sur des transcriptions automatiques. En effet, les données produites par le système de reconnaissance de la parole ne sont pas structurées en phrase ou en paragraphe (elles ne contiennent ni ponctuation, ni majuscule). Dans (Guinaudeau, 2011) et (Claveau et Lefèvre, 2011) les auteurs utilisent le *groupe de souffle* de parole comme *unité de base*. Un groupe de souffle correspond à la parole prononcée par un locuteur entre deux respirations (pauses silencieuses).

1.3.3 Mesures de similarité

Une mesure de similarité permet d'estimer la proximité ou l'éloignement de vecteurs de même type. En segmentation thématique, la similarité est calculée entre les différentes parties de l'émission (représentées sous forme vectorielle) pour juger de leurs ressemblance ou dissemblance. Une faible valeur de similarité signifie que les deux vecteurs partagent très peu de mots en commun et donc traitent deux thèmes différents. À l'opposé, une valeur élevée correspond à des vecteurs contenant plusieurs mots en commun et donc traitant la même thématique. Les méthodes les plus connues dans l'état de l'art sont : *Jaccard* (Jaccard, 1901), *Cosine* (Salton *et al.*, 1975), *LIN* (Lin, 1998) et *Extended_*JACCARD (Curran et Moens, 2002).
La mesure *Jaccard* (notée dans ce manuscrit par Set_Jaccard) calcule la similarité entre deux vecteurs A et S en faisant simplement le rapport entre le nombre de termes communs et le nombre de tous les termes apparaissant dans les deux vecteurs. Les mesures *Lin* (Lin, 1998) et *Extended_*JACCARD sont deux versions pondérées de la mesure *Jaccard*. Le tableau 1.3 donne la définition de ces mesures telles qu'elles sont présentées dans (Curran et Moens, 2002).

1.4 Méthodes de segmentation thématique

Dans cette section, nous donnons un panorama général des algorithmes les plus utilisés. Parmi les systèmes de segmentation mis en place, on distingue ceux fondés sur des méthodes à base de mots et les systèmes multimodaux. Les documents standards sont composés uniquement des mots. Par conséquent, une seule modalité est exploitable. Cependant, les documents audiovisuels possèdent de nombreuses caractéristiques permettant de trouver les frontières thématiques comme la répartition des locuteurs, les mots prononcés, les titres incrustés, *etc.* Néanmoins, pour avoir un système générique (*i.e.* indépendant de

Set_JACCARD	$\frac{\lvert A \cap S \rvert}{\lvert A \cup S \rvert}$
Extended_JACCARD	$\frac{\sum_{t \in A \cap S} w^A(t) * w^S(t)}{\lVert W^A \rVert_2^2 + \lVert W^S \rVert_2^2 - \sum_{t \in A \cap S} w^A(t) * w^S(t)}$
LIN	$\frac{\sum_{t \in A \cap S} w^A(t) + w^S(t)}{\sum_{t \in A \cup S} w^A(t) + w^S(t)}$
Cosine	$\frac{\sum_{t \in A \cap S} w^A(t) * w^S(t)}{\lVert W^A \rVert_2 * \lVert W^S \rVert_2}$

Tableau 1.3 – *Mesures de similarité utilisées où A et S sont deux documents. $w^A(t)$ (resp. $w^S(t)$) désigne le poids du terme t dans le document A (resp. S). $\left\lVert W^S \right\rVert$ correspond à la norme du vecteur de mots W^S.*

toute sorte d'information structurelle sur l'émission traitée), certains indices devront être privilégiés par rapport à d'autres, comme les mots prononcés durant l'émission, en partant sur le postulat que *certains mots sont spécifiques à un thème*.

1.4.1 Segmentation thématique basée sur les indices lexicaux

Les méthodes lexicales ont été initialement appliquées sur des documents écrits. Elles considèrent qu'un changement de vocabulaire remarquable est le signe d'une transition thématique.

Dans cette section, nous présentons uniquement les méthodes purement statistiques qui ne font pas appel à des ressources externes. Des méthodes statistiques intégrant des relations sémantiques entre les mots sont présentées en section 6.3.

L'algorithme TextTiling

TextTiling (Hearst, 1997) est l'un des premiers algorithmes de segmentation thématique. Il est à l'origine de plusieurs progrès réalisés dans ce domaine. Il repose sur l'analyse de la distribution des mots. À l'aide d'une fenêtre glissante, la similarité entre des blocs adjacents (un bloc est constitué de N phrases) est calculée tout au long de l'émission (voir la figure 1.3). Une valeur de similarité proche de 1.0 signifie que les deux blocs partagent un nombre important

de mots et donc appartiennent au même thème. Par contre, une valeur faible indique qu'il y a peu de mots en commun et donc qu'il y a une forte possibilité d'être en présence de deux thèmes différents.

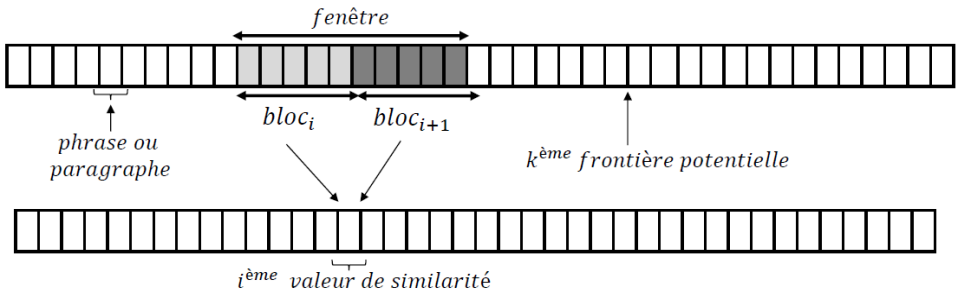


FIGURE 1.3 – *Calcul de la cohésion lexicale avec le principe de la fenêtre glissante.*

Il en résulte une courbe de cohésion lexicale (voir la figure 1.4) à partir de laquelle sont extraites les hypothèses de frontières. Dans cet algorithme, un minimum local est considéré comme une frontière potentielle. L'étape de détection

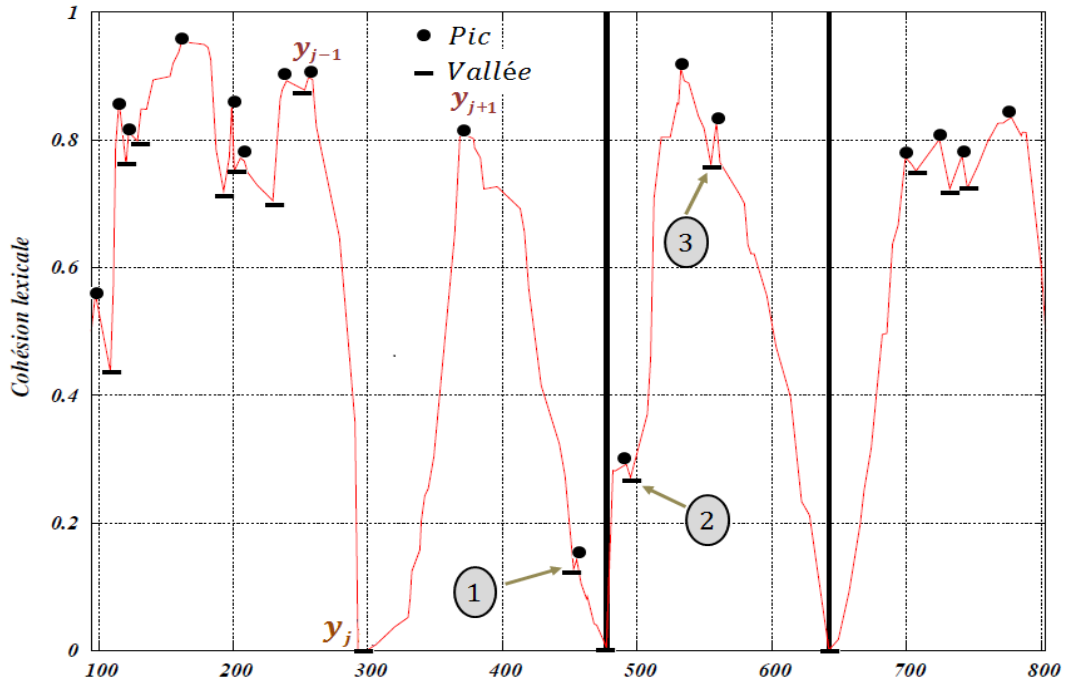


FIGURE 1.4 – *Courbe de la cohésion lexicale pour une émission TV.*

des frontières consiste d'abord à déterminer les pics et les vallée. Un pic correspond à un point de la courbe qui est entouré par deux valeurs plus faibles. Une vallée correspond à un point entouré par deux valeurs élevées. En d'autres termes, un pic indique des blocs fortement liés thématiquement, tandis qu'une vallée est susceptible de soulever une rupture thématique.

La profondeur de chaque vallée est calculée comme la somme des différences entre pics à gauche et à droite par rapport à la vallée en question. Pour une vallée v_j, la profondeur est donnée par

$$depth(v_j) = (y_{j-1} - y_j) + (y_{j+1} - y_j) \tag{1.1}$$

où y_c est la valeur de similarité à la position c

Pour déterminer le changement de thème à partir de la courbe de cohésion, il est possible que l'étape de détection des frontières soit précédée par un lissage. Ce lissage permet, dans certains cas, d'éliminer les petites vallées qui peuvent fausser le calcul des profondeurs (comme les vallées 1, 2 et 3 de la figure 1.4). Il est réalisé en déplaçant une fenêtre sur la courbe et en remplaçant la valeur de cohésion associée au centre de la fenêtre par la moyenne des valeurs de cohésion de toutes les positions incluses dans la fenêtre. Un lissage de taille n est défini par la moyenne des $n/2$ valeurs de similarité précédentes + la valeur de similarité courante + les $n/2$ valeurs de similarité suivantes.

Enfin, un seuil est appliqué sur la valeur de la profondeur de chaque vallée pour déterminer si elle correspond bien à une frontière ou non. Une vallée est considérée comme une rupture thématique si elle est inférieure à un seuil donné.

L'algorithme C99
L'algorithme *C99* a été proposé par (Choi, 2000). Il repose également sur le calcul de la cohésion lexicale. Après les pré-traitements standards, les mesures de similarité sont calculées pour n'importe quel couple de phrases, on obtient ainsi une matrice de similarité de taille $n \times n$ (où n est le nombre de phrases).

L'originalité de cet algorithme réside dans l'exploitation indirecte de la matrice de similarité à travers une matrice de rang. Celle-ci opère un classement local de chaque case de la matrice de similarité vis-à-vis des cases voisines au sein d'un masque de classement. Ce dernier n'est autre qu'une sous-matrice de taille paramétrable qui sert à effectuer le classement. Le rang se calcule selon l'équation 1.2

$$rang = \frac{Nombre\,d'éléments\,ayant\,une\,similarité\,inférieure\,dans\,le\,masque}{Nombre\,d'éléments\,réellement\,présents\,dans\,le\,masque} \tag{1.2}$$

La figure 1.5 est un exemple de calcul de la matrice de rang [4] avec un masque de taille 3×3. Les valeurs de similarité sont comprises entre 0 et 8. Dans l'étape 1, la case $(3, 2)$ correspondant au score 8. Elle est entourée par 7 valeurs de similarités inférieures. La case $(3, 2)$ prend donc la valeur 7 dans la matrice de rang.

4. Pour des raisons de visualisation, les valeurs de la matrice de rang n'ont pas été normalisées par le nombre d'éléments réellement présents dans le masque.

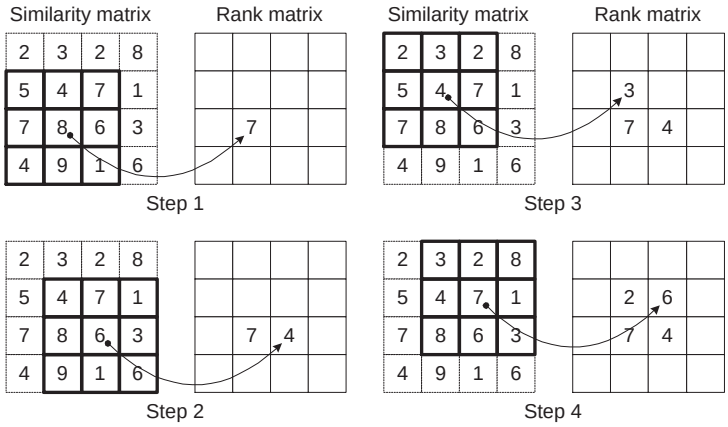


FIGURE 1.5 – *Calcul de la matrice de rang à partir de la matrice de similarité.*

Les segments thématiques sont ensuite identifiés *via* un processus de clustering inspiré de l'algorithme de Reynar (Reynar, 1994).

L'algorithme MinCut

Dans (Malioutov et Barzilay, 2006), le problème de la segmentation thématique est modélisé comme une tâche de partitionnement de graphe. Celui-ci est constitué à partir de l'ensemble des phrases (sommets du graphe) dans le texte et les poids des arêtes entre chaque paire de sommets $w(u, v)$ correspond à la valeur de similarité entre les deux phrases u et v. Il s'agit ensuite de minimiser la cohésion entre les différents segments thématiques du document. Les auteurs mettent en place une méthode qui consiste à découper le graphe en k classes, en minimisant le critère de la coupe. Les segments obtenus devront avoir un contenu homogène et être différents les uns des autres.

Construction du graphe

Avant de calculer les valeurs de similarité entre les phrases, des pré-traitements standards sont appliqués, suivis par une étape de pondération de termes. Les auteurs introduisent la pondération intra-document (le principe est détaillé dans le chapitre 4).
À l'issue de cette étape, un graphe non orienté $G = \{V, E\}$ est construit où V est l'ensemble de sommets correspondant à des phrases dans le texte et E est l'ensemble des poids des arêtes.

Détection des frontières

Nous considérons d'abord le problème de partition binaire d'un graphe en deux ensembles A et B. Le but est de minimiser le critère de la coupure, définie par $NCut(A, B)$. Elle correspond à la somme normalisée des coûts associés à chacun des arcs reliant les segments. $NCut(A, B)$ est donnée par :

$$Ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \tag{1.3}$$

où : $Cut(A,B) = \sum_{u \in A, v \in B} w(u,v)$

$$vol(A) = \sum_{u \in A, v \in V} w(u,v)$$

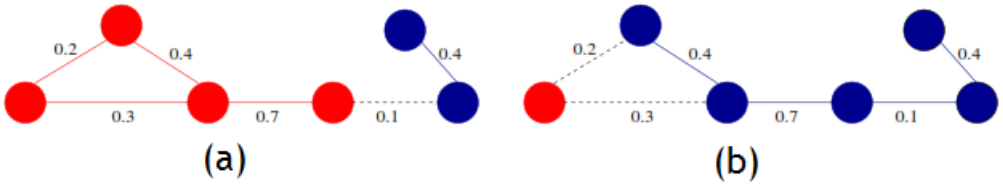


FIGURE *1.6* – *Exemple d'une coupure d'un graphe binaire.*

Dans les deux exemples de partitionnement illustrés dans la figure 1.6, les valeurs de critère de coupure se calculent ainsi :
Pour le graphe (a) :
$cut(A,B) = 0.1, vol(A) = 1.7, vol(B) = 0.5$ et $Ncut(A,B) = \frac{0.1}{1.7} + \frac{0.1}{0.5} = 0.26$
Pour le graphe (b) :
$cut(A,B) = 0.5, vol(A) = 2.1, vol(B) = 0.5$ et $Ncut(A,B) = \frac{0.5}{2.1} + \frac{0.5}{0.5} = 1.2$
La partition *(a)* prend la valeur la plus faible, c'est donc celle qui sera considérée comme étant la meilleure segmentation.

La formule 1.3 permet la découpe binaire d'un graphe binaire. Pour un graphe (V) contenant $A_1, .., A_k$ partitions, le découpage normalisé est donné par

$$Ncut_k(V) = \frac{cut(A_1, V - A_1)}{vol(A_1)} + ... + \frac{cut(A_k, V - A_k)}{vol(A_k)} \tag{1.4}$$

où $V - A_k$ est la différence entre la $k^{ème}$ partition et le graphe entier.
La minimisation normalisée du découpage d'un graphe est un problème NP-complet. Les auteurs utilisent alors la programmation dynamique pour que le découpage s'effectue dans un temps polynomial.

$$C[i,k] = \min_{j<k} \left[C[i-1,j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]} \right] \tag{1.5}$$

$$B[i,k] = \operatorname*{argmin}_{j<k} \left[C[i-1,j] + \frac{cut[A_{j,k}, V - A_{j,k}]}{vol[A_{j,k}]} \right] \tag{1.6}$$

où $C[i,k]$ est la valeur de coupure optimale pour les k premières phrases en i segments.
$A_{j,k}$ est l'ensemble de nœuds commençant par le $j^{ème}$ nœud et se terminant par le $k^{ème}$ nœud.
$B[i,k]$ est une table contenant la séquence optimale de la segmentation thématique.
La complexité de l'algorithme avec la programmation dynamique est $O(KN^2)$, où K est le nombre de partitions et N est le nombre de nœuds (phrases) dans le graphe.

L'algorithme WLL
L'algorithme *WLL* proposé par (Sitbon et Bellot, 2007) repose sur le principe des chaînes lexicales. Les frontières sont déterminées non seulement à partir de la répétition des mots mais aussi à partir de leur position. Une chaîne lexicale relie les mots ayant la même forme écrite. Elle commence à la première occurrence et se termine à la dernière occurrence de ce terme. Une chaîne est divisée en sous chaînes si la distance entre deux occurrences consécutives dépasse un certain seuil appelé *hiatus*. On peut alors recenser pour chaque phrase les chaînes actives. La figure 1.7 illustre le processus de construction des chaînes lexicales,

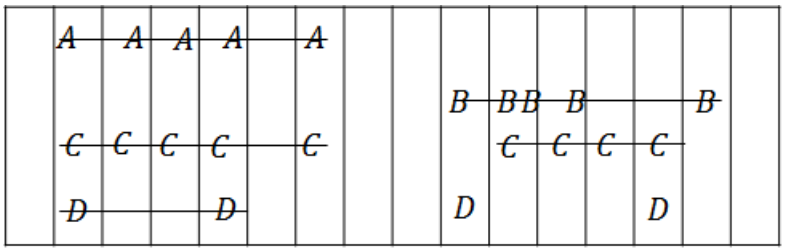


FIGURE 1.7 – *Construction des chaînes lexicales pour les termes A, B, C et D tout au long du document, avec un hiatus=3.*

avec un $hiatus = 3$. L'importance de l'utilisation du seuil *hiatus* est lisible dans la détermination du début et fin de chaque chaîne lexicale. Le principe de pondération proposé dans (Galley *et al.*, 2003) a été utilisé. Il induit que le poids d'une chaîne lexicale est calculé en fonction de sa taille, son nombre d'occurrences ainsi que la catégorie lexicale du terme considéré. Le poids d'une chaîne R_i est défini par :

$$w(R_i, t_i) = cat_{t_i} \times nb_{t_i} \times log\left(\frac{L}{L_i}\right) \tag{1.7}$$

où $R_1, R_2, ..., R_n$ est l'ensemble de toutes les chaînes du document correspondant aux termes $t_1, t_2, ..., t_n$ de longueur $L_1, L_2, ..., L_n$ respectivement ; cat_{t_i} est la catégorie syntaxique du terme t_i formant la chaîne R_i et L est la longueur du document.
Le seuil *hiatus* peut être empiriquement défini pour l'ensemble des chaînes,

comme il peut être estimé séparément.
La similarité entre les phrases adjacentes est calculée en utilisant le principe général de *TextTiling*. La similarité est calculée tout au long de l'émission par le biais d'une fenêtre glissante de taille n). La similarité est alors donnée par :

$$cosine(A, B) = \frac{\sum_i w(A, t_i) \times w(B, t_i)}{\sqrt{\sum_i w^2(A, t_i) \times \sum_i w^2(B, t_i)}} \tag{1.8}$$

où A et B sont les ensembles des vecteurs représentant les poids des chaînes lexicales actives dans les n phrases avant et après.

L'algorithme Utiyama
Les auteurs de (Utiyama et Isahara, 2001) proposent un modèle statistique pour trouver la segmentation la plus cohérente possible dans un document W composé de n phrases réparties en m segments. La probabilité d'une segmentation S est définie par la règle de Bayes :

$$P(S|W) = \frac{P(W|S)P(S)}{P(W)} \tag{1.9}$$

L'objectif est de trouver la meilleure segmentation parmi toutes les segmentations possibles, c'est à dire celle qui maximise le numérateur de l'équation 1.9. Notons que la probabilité $P(W)$ est constante pour le texte W.

$$\hat{S} = argmax\ P(W|S)P(S). \tag{1.10}$$

Les auteurs supposent que chaque thème a une distribution de mots différente et que les m segments thématiques sont indépendants. Soit
n_i le nombre de mots du S_i et $W = W_1, W_2, ..., W_m$
où W_i est l'ensemble de mots du $i^{ème}$ segment ($W_i = w_1^i, .., w_{n_i}^i$)
w_j^i est le $j^{ème}$ mot du segment S_i.
La probabilité $P(W|S)$ se calcule alors ainsi :

$$\begin{aligned} P(W|S) &= P(W_1, W_2, ..., W_m|S) \\ &= \prod_{i=1}^{m} P(W_i|S_i) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} P(w_j^i|S_i) \end{aligned}$$

La probabilité $P(w_j^i|S_i)$ est définie par

$$P(w_j^i|S_i) \equiv \frac{f_i(w_j^i) + 1}{n_i + k} \tag{1.11}$$

où k est le nombre de mots différents dans l'émission.

$f_i(w_j^i)$ est le nombre d'occurrences du mot w_j^i dans W_i. Formellement $f_i(w_j^i)$ s'écrit de la manière suivante

$$f_i(w_j^i) \equiv \sum_{l=1}^{n_i} \nu(w_l^i, w_j^i) \tag{1.12}$$

où : $\nu(w_k^i, w_j^i) = 1$ si w_k^i et w_j^i sont les mêmes mots sinon $\nu(w_k^i, w_j^i)$ vaut 0.

Les auteurs supposent que $P(S) \equiv h^{-m}$ où h est le nombre de mots dans le document et m le nombre de segments.
Maximiser l'équation 1.10 revient à minimiser le coût de segmentation :

$$\widehat{S} = \underset{S}{arg\,max}\ P(W|S)P(S) = \underset{S}{arg\,min}\ C(S) \tag{1.13}$$

$$\begin{aligned} \text{où C(S)} \quad &\equiv -\log P(W|S)P(S) \\ &= \text{-} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \frac{f_i(w_j^i) + 1}{n_i + k} + m \log n \end{aligned}$$

Le document à segmenter $(W = w_1, w_2, ..., w_n)$ est représenté sous forme de graphe $G = (V, E)$

où $V = \{g_i \mid 0 \leq i \leq n\}$ correspond aux sommets (frontières potentielles). g_i est positionnée entre le mot w_i et w_{i+1}. Donc g_0 est placée avant w_1 et g_n après w_n.
$E = \{e_{i,j} \mid 0 \leq i \leq j \leq n\}$ où $e_{i,j}$ représente le poids du segment allant du mot w_i à w_j

L'algorithme de programmation dynamique est utilisé pour trouver les points de rupture (*i.e* le chemin de coût minimal).

(Guinaudeau et Hirschberg, 2011) ont adapté l'algorithme de (Utiyama et Isahara, 2001) aux spécificités des documents oraux (erreurs de transcription automatique, manque de répétition dans la parole préparée, *etc.*). Pour cela, les auteurs exploitent les mesures de confiance associées à chacun des mots de la transcription automatique ainsi que l'utilisation de relations sémantiques. Dans (Simon *et al.*, 2013) , les auteurs ont proposé une méthode de segmentation thématique basée sur l'algorithme de (Utiyama et Isahara, 2001) qui combine la cohésion lexicale et la rupture lexicale, identifiant des zones de continuités et de ruptures dans l'organisation globale des données

1.4.2 Segmentation thématique multimodale

Un journal d'information télévisuel contient des indices de nature diverse qui peuvent être utiles pour la détection des changements de thèmes. Les systèmes de segmentation multimodaux utilisent des indices provenant d'au moins

deux sources d'informations. La fusion de différentes modalités se fait essentiellement dans un cadre supervisé (*i.e* à l'aide d'un classifieur sur des données annotées). Par exemple, (Xie *et al.*, 2010a) mettent en place un modèle d'*entropie maximale* pour combiner les différentes sources d'information. Le classifieur *CRF* (Conditional Random Fields) a été utilisé pour rassembler toutes les sources d'informations (Wang *et al.*, 2012) et (Bouchekif *et al.*, 2013a). D'autres algorithmes d'apprentissage comme les machines à vecteurs supports (Hsu *et al.*, 2005), (Georgescul *et al.*, 2006b) ou encore les modèles de Markov cachés (Tür *et al.*, 2001a) ont aussi été utilisés. Les arbres de décision ont été largement utilisés dans la segmentation thématique pour combiner différentes sources d'information. On peut citer les travaux de (Tür *et al.*, 2001b), (Beeferman *et al.*, 1999) et (Passonneau et Litman, 1997).
Plusieurs systèmes (Dumont et Quénot, 2012), (Chaisorn *et al.*, 2003), (Wang *et al.*, 2012), *etc.* fusionnent trois sources d'indices : lexicaux, acoustiques et visuels.

Indices acoustiques

— *Durée de la pause (silence ou musique) :* un nouveau thème est généralement précédé par une pause. Par ailleurs, le passage à d'autres informations comme la météo et le sport est souvent précédé d'un jingle. Cet indice a été utilisé dans plusieurs travaux comme (Wang *et al.*, 2012) et (Shriberg *et al.*, 2000).

— *Environnement acoustique :* un nouveau thème démarre généralement dans le studio et très rarement dans un milieu bruité. Dans (Wang *et al.*, 2012), les auteurs donnent six valeurs à l'attribut : purement parole, parole bruitée, parole accompagnée de musique, bruit, musique et silence.

— *Changement de locuteur :* un journal d'information contient plusieurs locuteurs (présentateur principal, reporters, invités, intervenants externes, *etc.*). L'exploitation de cette information s'avère également un bon indicateur pour repérer les frontières. (Wang *et al.*, 2012) considèrent qu'un changement de locuteur peut coïncider avec un changement de thème. Ils représentent le changement du locuteur par un attribut binaire. (Dumont et Quénot, 2012) gardent l'enchaînement de locuteurs en associant à chaque groupe de souffle l'indice du locuteur correspondant.

— *Présentateur principal :* selon plusieurs travaux, cet indicateur est considéré comme le plus robuste pour repérer les changements de thèmes à condition que l'émission soit présentée par un unique présentateur principal. Ce dernier est chargé d'assurer les transitions entre les thèmes. Cependant, cet indice peut générer du bruit si l'émission ne dispose pas de présentateur, ou bien si l'émission contient deux ou plusieurs présentateurs.

Indices visuels

— *Titre incrusté :* un nouveau thème peut être accompagné par un titre indiquant l'objet de la nouvelle section thématique. (Wang *et al.*, 2012) mesurent la différence de durée entre les apparitions du titre et de la frontière potentielle [5], la valeur considérée est celle de l'attribut.

— *Visage du présentateur principal :* généralement le lancement d'un nouveau thème se fait dans un studio par le présentateur principal de l'émission.

— *Logo :* chaque chaîne possède un logo qui prend la même position et de façon continue, sauf pendant les publicités où il sera absent. À partir de ce constat (Dumont et Quénot, 2012) donnent une valeur binaire à l'attribut logo (1 si le logo est présent et 0 sinon).

La combinaison de ces indices est en règle générale profitable à la tâche de segmentation thématique ((Wang *et al.*, 2012), (Dumont et Quénot, 2012)). Toutefois, certains indices sont fortement liés aux règles éditoriales de chaque chaîne télévisée, ce qui nécessite des informations *a priori* sur les émissions à segmenter. Par exemple, le système développé dans (Rosenberg et Hirschberg, 2006) nécessite des informations sur la provenance de la source des émissions, par exemple CNN, ABC, *etc.*, pour construire les différents classifieurs de chaque émission. Le titre incrusté peut être un très bon indicateur d'un changement de thème, mais les algorithmes basés sur cet attribut échouent s'ils n'existent pas et de façon similaire pour les émissions tournées hors studio. Les indices visuels, quant à eux, rendent la solution très spécifique aux données étudiées (Wang *et al.*, 2012), (Dumont et Quénot, 2012), *etc.* De plus, (Dumont et Quénot, 2012) indiquent qu'en terme de performance, les indices visuels sont moins importants que les indices issus de l'audio.

1.5 Les corpus de la segmentation thématique

Pour évaluer la tâche de la segmentation thématique, il est nécessaire de disposer de documents (textuels ou audiovisuels) annotés. Le choix du corpus est généralement dicté par l'application visée et le coût de l'annotation. Celle-ci peut s'effectuer manuellement ou automatiquement. À noter que la plupart des travaux dédiés à la segmentation thématique utilisent leurs propres corpus.

Le corpus *C99* conçu par (Choi, 2000) est considéré comme étant le plus utilisé dans l'état de l'art (Misra *et al.*, 2009), (Galley *et al.*, 2003), *etc.* Il est composé de 7000 articles sélectionnés aléatoirement à partir du corpus *Brown* [6] de langue anglaise. Ces articles sont regroupés en 700 documents, où chaque document est la concaténation de 10 segments thématiques. Un segment est composé des

5. Les auteurs considèrent le silence ou la musique comme des frontières potentielles.
6. https ://en.wikipedia.org/wiki/Brown_Corpus

n premières phrases d'un article. Quatre catégories de segments ont été considérées : $3-11$, $3-5$, $6-8$ et $9-11$. Un document qui fait partie de la catégorie $3-5$ signifie qu'il contient uniquement les segments dont le nombre de phrases est compris entre 3 et 5. L'avantage de ce corpus est qu'il ne pose aucun problème d'annotation, puisque la fin de chaque segment est une frontière thématique. La figure 1.4 donne une description du corpus.

nombre de phrases	$3-11$	$3-5$	$6-8$	$9-11$
nb. documents	400	100	100	100

Tableau 1.4 *– Corpus de Choi - C99*

Le corpus ICSI (Janin *et al.*, 2003) est largement utilisé dans le domaine du traitement automatique de la parole. Il contient 75 documents transcrits automatiquement issus d'enregistrements de réunions [7] (d'environ une heure chacune). (Shriberg *et al.*, 2004) mettent à disposition la segmentation thématique de référence [8]. Ce corpus a été exploité dans plusieurs travaux comme (Eisenstein et Barzilay, 2008) et (Galley *et al.*, 2003).

Le corpus payant *TDT* (Topic Detection and Tracking) est devenu un standard dans la segmentation thématique. Ce corpus contient des documents en anglais, arabe et chinois (mandarin). Le corpus avec ses différentes versions (allant de TDT1 à TDT5) est utilisé pour évaluer de nombreux systèmes de segmentation comme (Rosenberg et Hirschberg, 2006) et (Xie *et al.*, 2010b).

Dans les travaux de (Guinaudeau, 2011), trois corpus ont été utilisés pour évaluer les méthodes de structuration automatique développées.

— Le premier est composé de 56 journaux télévisés diffusés en Février et Mars 2007 sur la chaîne de télévision *France 2*.

— Le deuxième contient 16 numéros de l'émission *Sept à Huit* diffusés en 2008 et 2009 sur *TF1*.

— Le troisième compte 7 numéros de l'émission *Envoyé Spécial* programmée entre 2008 et 2009 sur France 2.

Le premier corpus contient 1203 segments thématiques. Les deux autres sont beaucoup plus petits respectivement avec 86 et 26 segments pour les corpus *Sept à Huit* et *Envoyé Spécial*.
D'autres corpus sont également créés pour évaluer les systèmes de segmentation. On peut notamment citer (Malioutov et Barzilay, 2006) qui a créé son propre corpus à partir d'enregistrements de cours de physique.

7. Les fichiers audios et la transcription automatique sont disponibles sur la page http ://www1.icsi.berkeley.edu/Speech/mr/

8. Disponible sur le site http ://www1.icsi.berkeley.edu/ ees/dadb/

(Eisenstein et Barzilay, 2008) a mis à la disposition [9] un corpus issu d'un livre médical où chaque section est considérée comme un nouveau segment thématique.

1.6 Conclusion

Dans ce chapitre, nous avons présenté plusieurs algorithmes de segmentation thématique classiques. Nous avons vu que plusieurs indices peuvent être exploités pour segmenter un document multi-thèmes. Les indices utilisés dans les journaux d'information télévisuels peuvent être divisés en deux catégories : indices génériques et indices spécifiques. La première catégorie est applicable sur n'importe quel type d'émission. Elle concerne plus particulièrement les indices autour de la distribution des mots (répétition, mots de repérage, *etc.*). La deuxième catégorie dépend essentiellement des règles éditoriales de chaque émission (titre incrusté, visage du présentateur, *etc.*).
Les corpus utilisés dans les évaluations des systèmes de l'état de l'art nous ont mis en difficultés pour dégager des conclusions claires. Ce fut notamment le cas sur le positionnement de chaque approche par rapport aux autres en terme de performance. Cela est essentiellement dû à l'absence de campagnes d'évaluation dans le domaine de la segmentation thématique, ainsi qu'à l'utilisation de corpus non disponibles au public ou payants.
Après description des algorithmes, nous constatons qu'il existe des pistes inexplorées dans la segmentation des journaux télévisés. Les travaux présentés dans le chapitre 4 visent à donner un poids à chaque mot selon son degré d'importance. Ceci permet par exemple de pénaliser les mots apparaissant tout au long de l'émission et de favoriser les mots importants dans chaque thème. Nous proposons dans le chapitre 5 d'intégrer la distribution des locuteurs dans le calcul de la cohésion. Nous enrichissons la distribution des mots par des relations sémantiques, ces travaux sont présentés dans le chapitre 6. Nos contributions, qui concernent plus particulièrement l'enrichissement de la représentation vectorielle de l'émission, seront enfin présentées.

9. http ://groups.csail.mit.edu/rbg/code/bayesseg/

Chapitre 2

Protocole expérimental

Sommaire

2.1 Introduction

Pour attirer plus de téléspectateurs, chaque chaîne essaie de mettre en place des règles éditoriales plus innovantes tout en facilitant la transmission de l'information. Cela se traduit par une diversité dans les formats d'émissions qui compliquent la tâche des systèmes de segmentation thématique. Nous proposons dans ce chapitre une analyse détaillée des émissions constituant nos corpus. Celle-ci permettra également de mettre en œuvre un système de segmentation thématique générique capable de s'adapter à la diversité des formats. Par ailleurs, cela nous aidera à mieux interpréter les résultats.
Dans ce chapitre, nous décrivons puis analysons les trois corpus servant à tester et valider nos contributions. Ensuite, nous décrivons les métriques d'évaluation les plus utilisées dans l'état de l'art ainsi que celles que nous proposons.

2.2 Description générale des corpus

2.2.1 Construction de corpus

Les expériences ont été réalisées à partir des émissions de journaux télévisés (JT). Trois corpus ont été constitués tout au long de cette thèse :

— Le premier, nommé *MC-0813* [1], est utilisé pour régler les paramètres de notre algorithme de segmentation thématique. Il est constitué de 56 journaux télévisés issus de 8 chaînes françaises. Les émissions ont été collectées en deux périodes, la première concerne 7 chaînes : TF1, France2, France3, LCI, France24, Arte, M6 pour des émissions diffusées en 2008/2009, la seconde période concerne la chaîne D8 avec 23 JT datant d'octobre 2013.

— Le deuxième corpus, noté *MCS7-14* [2], est composé de 86 journaux télévisés collectés durant la période du 10 au 16 février 2014 en provenance de 8 chaînes françaises (TF1, France2, france3, M6, Arte, D8, NT1, Euronews). Ce corpus est exploité pour fixer les différents paramètres liés à l'intégration de relations sémantiques (chapitre 6) ainsi que ceux de l'algorithme de titrage (chapitre 7).

— Le troisième corpus *MCS5-15* [3] est composé de 26 journaux de 7 chaînes télévisées (TF1, France2, France3, M6, Arte, NT1, Euronews). Ces émissions ont été diffusées les 26 et 27 janvier 2015. Ce corpus est employé pour

1. MC désigne "multi-cannel", 0813 pour référencer à la période 2008-2013.
2. MC désigne "multi-cannel", S7-14 pour référencer la semaine 7 de 2014.
3. MC désigne "multi-cannel", S5-15 pour référencer la semaine 5 de 2015.

tester l'ensemble de la chaîne de traitements (segmentation thématique et titrage).

Des informations sur le nombre de frontières et sur la durée des thèmes sont données plus loin dans le tableau 2.1.

2.2.2 Transcription automatique

Pour traiter le contenu linguistique des documents oraux, le signal audio de chaque émission est transformé en texte correspondant aux paroles prononcées. Cette opération est réalisée par le biais d'un système de reconnaissance de la parole. Les systèmes de transcription ont montré leur efficacité sur les journaux télévisés. Cela tient essentiellement à leur nature :

— Un journal d'information contient majoritairement de la parole préparée. En effet, chaque journaliste, prépare son discours, ce qui lui permet d'éviter les disfluences durant ses interventions.

— Un journal d'information est majoritairement enregistré dans un milieu non bruité.

— La plupart des systèmes de transcription automatique utilisent des données d'apprentissage adéquates aux journaux d'information.

Dans nos expériences, nous utilisons l'outil *Voxsigma* de *Vocapia Research* basé sur le système du LIMSI (Gauvain *et al.*, 2002). Le taux d'erreurs mots sur 33 émissions du corpus *MC-0813* est de l'ordre 16.1%. En revanche, nous ne possédons pas de transcriptions manuelles pour les corpus *MCS7-14* et *MCS5-15*. Par conséquent, nous ne connaissons pas le taux d'erreur sur ces corpus.
Les débuts et fins de groupes de souffle utilisés pour déterminer les unités de base (voir la section 1.3.2) sont déterminés par le système de transcription automatique. Leur durée moyenne est de l'ordre de 4.7 secondes pour les corpus *MC-0813 (min=0.6, max=37.8)* et *MCS5-15 (min=0.8, max=35.4)*, et de 4.5 secondes pour le corpus *MCS7-14 (min=0.8, max=39.1)*.

2.2.3 Segmentation thématique de référence

L'annotation humaine est très coûteuse à la fois en terme d'argent et de temps. D'ailleurs, il très difficile de trouver gratuitement un corpus annoté manuellement. En effet, la plupart des ressources disponibles qui disposent d'une segmentation thématique sont construites de façon automatique en concaténant plusieurs documents, chacun de ces derniers devant traiter une seule thématique. Ce type de corpus a été mis en œuvre par (Choi, 2000) et a été utilisé dans plusieurs travaux. Cependant, dans certains travaux, des corpus privés ont été

construits. Le choix du type de documents est alors dicté par l'application visée.
De notre côté, nous avons fait appel à un annotateur pour produire une segmentation de référence tout en respectant la définition d'un thème que nous avons adoptée (voir la partie 1.2). Plus précisément, la tâche consiste à poser une limite entre la fin d'un thème et le début d'un autre. Pour une meilleure lisibilité du contenu de l'émission, une petite description décrivant le contenu de chaque segment a été insérée. L'annotation des trois corpus a été réalisée *via* l'outil *Transcriber* [4].

Il faut noter que nous procédons en écartant certaines parties spécifiques de l'émission : il s'agit principalement de la première et de la dernière partie du journal car elles ne contiennent que les grands titres et leur rappel. Par

FIGURE 2.1 – *Rappel des titres au milieu du journal*

ailleurs, dans certains journaux (en particulier les journaux d'une plus grande longueur), le présentateur donne de petites descriptions au milieu du journal (voir la figure 2.1) sur les thèmes qui seront traités par la suite (titres intermédiaires). Ces parties perturbent notre algorithme de segmentation thématique basé sur la répétition de mots ainsi que l'algorithme d'évaluation. Nous considérons que la détection de ces zones aurait pu être effectuée automatiquement en utilisant un outil de détection de jingles par exemple. Dans le cas où l'émission contient des titres intermédiaires, elle est subdivisée en parties qui seront traitées séparément (voir la figure 2.2). Le découpage en parties est hétérogène puisque certaines émissions ne sont pas concernées par la subdivision tandis que d'autres contiennent une ou deux zones de titres intermédiaires.

4. http ://trans.sourceforge.net/en/presentation.php

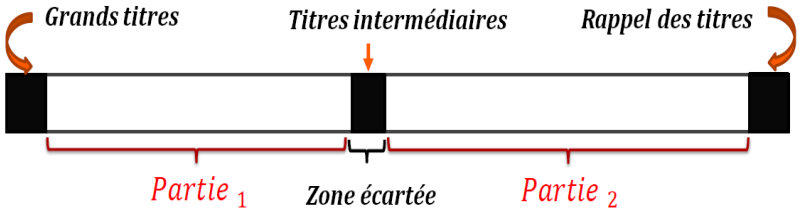


FIGURE 2.2 – *Traitement séparé pour chaque partie de l'émission.*

2.3 Analyse des corpus

2.3.1 Répartition des chaînes

La figure 2.3 représente la répartition des chaînes dans chaque corpus. Durant la période de la collecte des deux corpus *MCS7-14* et *MCS5-15*, nous avons essayé de récupérer tous les journaux télévisés sans aucune préférence de chaînes. La différence de distribution entre les chaînes au sein d'un même corpus, re-

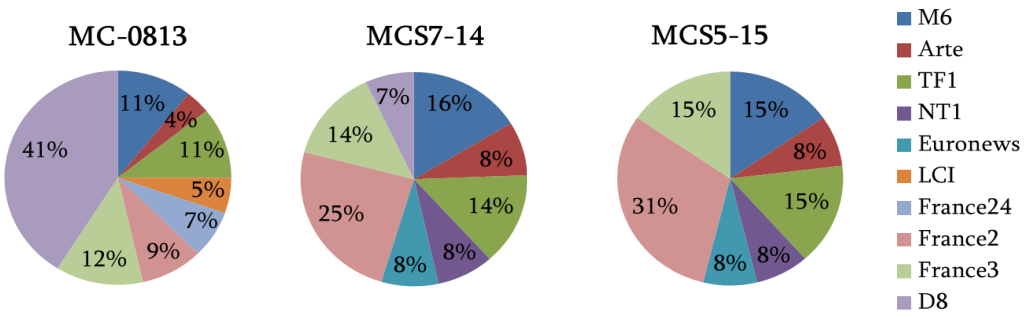


FIGURE 2.3 – *Répartition des chaînes en terme de nombre d'émissions dans les trois corpus.*

vient essentiellement au nombre d'éditions diffusées par chaque chaîne. Par exemple, France2 possède 4 éditions : le journal de 7 Heures, celui de 8 Heures, ceux de 13 Heures et 20 Heures, ce qui augmente la présence de la chaîne dans les corpus. À noter que chaque édition a ses caractéristiques spécifiques (la durée de l'émission, la durée des thèmes, le nombre de présentateurs, *etc.*) y compris au sein d'une même chaîne.
La figure 2.4 donne la composition détaillée des corpus pour les chaînes possédant plusieurs éditions. Pour les autres chaînes (NT1, LCI, D8, Euronews, France24) une seule édition par jour a été retenue.

Traiter des émissions provenant de plusieurs chaînes collectées sur une courte période entraîne quelques caractéristiques particulières. Nos corpus sont marqués par :

FIGURE 2.4 – *Composition des trois corpus par émission.*

La variété et la richesse : les émissions sont issues de 10 chaînes françaises traitant de domaines variés comme la politique, le sport, le cinéma, *etc.* De plus, chaque émission possède ses propres règles éditoriales. Celles-ci peuvent être séparées en deux catégories : *traditionnelle* et *moderne*.
Dans les émissions dites *traditionnelles*, les thèmes sont majoritairement lancés en plateau par le présentateur principal. Le lancement est suivi d'un reportage ou d'une interview. Certains thèmes sont intercalés par des brèves lues par le présentateur principal ou par un autre journaliste sans qu'un reportage ne vienne illustrer ses propos (c'est le cas du journal d'Arte par exemple).
Les émissions de type *moderne* ont des règles éditoriales un peu différentes. En effet, il y a des émissions qui n'ont pas de présentateur principal. C'est le cas d'Euronews et du journal de France2 de 7H ayant plus d'un présentateur. Certains JT contiennent une succession de reportages, sans retour au plateau et sans être introduits par le présentateur principal (c'est le cas du journal du soir de France 3 incluant en fin de programme une succession de reportages issus des éditions régionales). La dernière colonne du tableau 2.2 indique le type de chaque émission (*i.e.*traditionnel (T) ou moderne (M)).

La dynamicité : beaucoup d'émissions de nos corpus traitent d'informations qui évoluent avec le temps (évènements chronologiques). Prenons le cas des éditions *13 Heures* et *20 Heures* du journal de France2 qui ont été diffusées le 13/02/2015. Les deux éditions traitent des JO de Sotchi. Dans celle de *13 Heures*, il est question de la déception côté français, et dans l'édition de *20 Heures*, le journal s'intéresse à la deuxième médaille d'or pour Martin Fourcade.

2.3.2 Analyse des segments thématiques

Un journal télévisé est constitué d'un ensemble de thèmes présentés sous diverses formes : reportages, débats, interviews ou brèves. La durée des thèmes

peut aller de quelques secondes à des dizaines de minutes.

Le tableau 2.1 illustre les caractéristiques de notre corpus après la suppression des parties perturbatrices. Le corpus *MC-0813* contient 510 frontières thématiques correspondant à 570 segments d'une durée moyenne de 106*s*.

Corpus	MC-0813	MCS7-14	MCS5-15
Durée du corpus	16.5*h*	23.3*h*	9.9*h*
Nbre frontières	510	895	271
Nbre segments	570	997	297
Durée moy. (min, max)	106.0 (8.6,518.8)	105.1 (5.2,1145.4)	120.5 (5.1,655.5)

Tableau 2.1 *– Description des corpus*

Le corpus *MCS7-14* possède 895 frontières équivalentes à 997 segments, la taille des segments peut aller de 5.2*s* à 1145.4*s* ($\simeq$ 19*min*) avec une durée moyenne de 105*s*. Le segment le plus long de ces trois corpus [5] est considéré comme un cas particulier, puisque il est le seul qui dépasse 11 minutes. Le corpus *MCS5-15* est beaucoup plus petit, il est constitué de 297 segments d'une durée moyenne de 120*s*.

Le tableau 2.2 met en évidence la variabilité de la taille des segments entre les différents types d'émissions mais aussi au sein d'une même émission. Les thèmes traités peuvent être accompagnés par des reportages ou des interviews. Les thèmes par conséquent sont plus longs. Le présentateur peut également résumer lui-même en quelques secondes une information (brève). Les segments sont donc courts. Ainsi, certaines émissions, comme par exemple 7 heures ou 8 heure, ou NT1 ont une durée moyenne de segments courts plus courte que d'autres émissions comme D8, le 13 heures de France 2.

Nous souhaitons différencier les segments en fonction de leur taille. Par la suite, nous considérons les segments ayant une durée supérieure à 30*s* comme *longs*, les segments en déjà 30*s* sont considérés comme courts. Parmi les 570 segments du corpus MC-0813, 105 (18.4%) sont de type *courts* tandis que les corpus *MCS7-14* et *MCS5-15* en contiennent respectivement 236 (23.7%) et 70 (23.6%)b. La figure 2.5 donne la répartition des segments pour chaque type d'émission selon leur nature (long ou court). De la lecture de la figure 2.5, il ressort que toutes les émissions contiennent les deux types de segments avec une large domination des longs, en particulier pour les JTs de D8.

5. Le thème le plus long appartient au journal de TF1 de 13 heures du 16 février 2014 http ://lci.tf1.fr/jt-we/videos/2014/zoom-sur-la-cite-majestueuse-de-saint-petersbourg-8366074.html

Émission	Durée moyenne des segments (min ; max)			Type
	MC-0813	MCS7-14	MCS5-15	
Arte_LeJournal	84.4 (19.8 ; 321.8)	123.6 (19.7 ; 341.4)	158.9 (14.0 ; 452.1)	T
D8_LeJournal	107.0 (18.8 ; 273.6)	170.0 (39.9 ; 258.2)	-	T
Euronews_LeJournal	-	68.8 (15.5 ; 159.2)	67.6 (19.0 ; 314.4)	M
France24	96.5 (25.1 ; 208.6)	-	-	T
France2_7Heures	-	62.7 (9.5 ; 290.0)	59.8 (13.7 ; 152.4)	M
France2_8Heures	-	56.5 (9.6 ; 137.4)	61.6 (12.8 ; 151.6)	T
France2_13Heures	151.4 (16.6 ; 323.2)	147.4 (8.7 ; 447.3)	190.3 (10.8 ; 560.7)	T
France2_20Heures	113.6 (11.3 ; 275.6)	126.2 (9.9 ; 447.5)	176.5 (11.7 ; 655.6)	T
France3_12/13Heures	108.3 (14.7 ; 272.5)	84.9 (11.4 ; 274.2)	118.4 (15.2 ; 320.9)	T
France3 _19/20Heures	132.60 (15.6 ; 374.1)	101.2 (13.8 ; 258.2)	132.7 (18.0 ; 315.2)	T
LCI_LeJournal	42.7 (8.6 ; 192.0)	-	-	T
M6_12h45	83.4 (12.1 ; 276.6)	81.7 (8.8 ; 232.9)	105.9 (17.3 ; 300.0)	T
M6_19h45	-	92.0 (20.7 ; 373.7)	117.6 (18.8 ; 360.0)	T
M6_SixMinutes	85.3 (22.9 ; 205.5)	-	-	M
TF1_13Heures	112.4 (10.5 ; 458.2)	126.5 (5 .2 ; 1145.4)	113.1 (5.1 ; 265.2)	T
TF1_20Heures	107.1 (9.3 ; 448.8)	121.8 (15.3 ; 451.7)	131.3 (14.1 ; 348.9)	T
NT1_LeJournal	-	54.6 (15.9 ; 107.0)	68.2 (21.0 ; 83.2)	M

Tableau 2.2 – *Description des émissions en terme de longueurs de segments (les émissions sont classées par ordre alphabétique).* ***T*** *: émission de type traditionnel et* ***M*** *: émission de type moderne.*

2.3.3 Redondance des mots

La répétition des mots dans chaque thème est un avantage pour les systèmes de segmentation se basant uniquement sur les mots prononcés durant l'émission. Il est alors intéressant de connaître la redondance moyenne des mots de chaque segment. Pour cela, nous avons mis en œuvre les étapes suivantes :

— *Étape 1* : calcul de la fréquence des mots dans chaque segment

— *Étape 2* : tous les mots de l'émission ayant la même fréquence sont regroupés ensemble.

— *Étape 3* : comptage du nombre de mots dans chaque ensemble.

Le tableau 2.3 illustre un exemple simple de calcul de la répétition des mots dans une émission contenant deux segments. La fréquence des mots dans chaque segment est représentée par la troisième colonne du tableau. Les mots ayant le même nombre d'occurrences sont regroupés. La troisième étape donne :

— Huit mots ont une fréquence de 1 (*europe, bruxelles, médecin, parlement, autoroute, lundi, périphérique, galère*), ce qui représente 42.1% des mots.

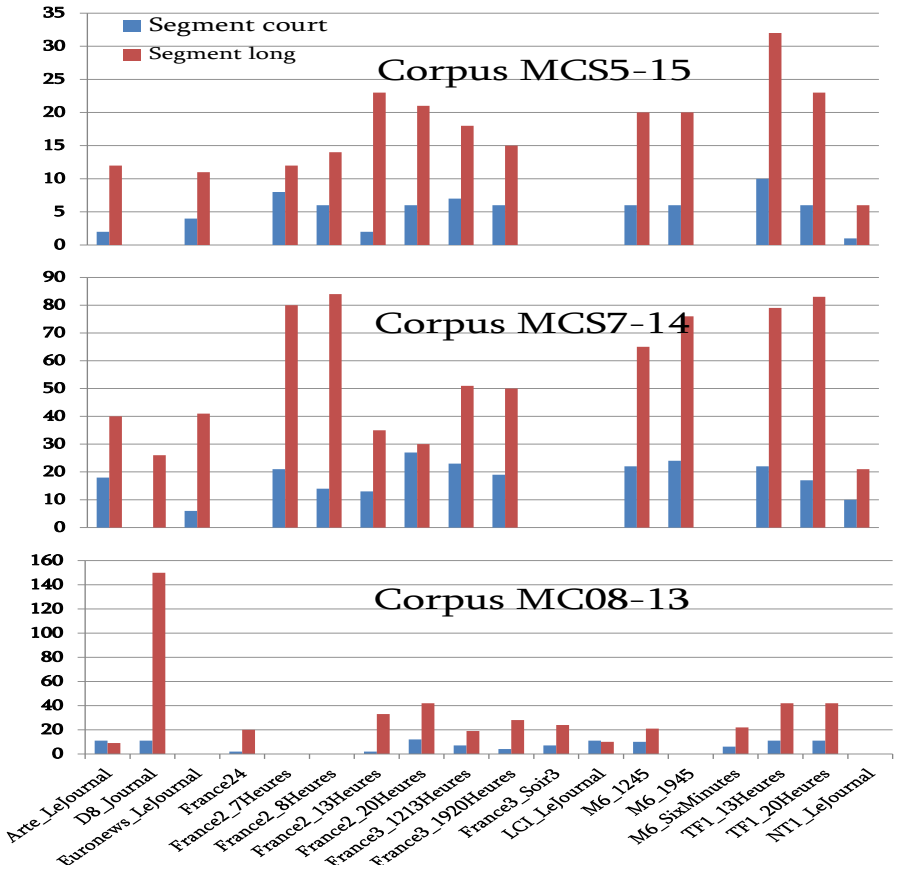


FIGURE 2.5 – *Description des segments courts et longs pour chaque émission dans les trois corpus.*

thème	Mots	Fréquence
1	taxi, vtc	5
	grève, conducteur	3
	paris, aéroport	2
	autoroute, lundi, périphérique, galère	1
2	suisse, immigration	4
	référendum, contre	3
	étranger	2
	europe, bruxelles medecin, parlement	1

Tableau 2.3 – *La redondance des mots au niveau des segments pour un document composé de deux thèmes.*

— Quatre mots ont une fréquence de 3 (*grève, conducteur, référendum, contre*),

ce qui représente 21.1% des mots.

— Trois mots ont une fréquence de 2 (*paris, aéroport, étranger*), ce qui représente 15.8% des mots.

— Deux mots ont une fréquence de 4 (*suisse, immigration*), ce qui représente 10.5% des mots.

— Deux mots ont une fréquence de 5 (*taxi, vtc*), ce qui représente 10.5% des mots.

Le même principe de calcul présenté dans l'exemple précédent a été appliqué sur les trois corpus considérés, les résultats obtenus sont donnés dans le tableau 2.4.

	1	2	3	4	≥ 5
MC-0813	82.7%	11.9%	3.0%	1.2%	1.2%
MCS7-14	83.6%	11.1%	3.1%	1.1%	1.1%
MCS5-15	82.1%	11.8%	3.4%	1.2%	1.5%

***Tableau 2.4** – La répétition moyenne des mots dans un segment pour les trois corpus.*

Il découle que nos données souffrent d'un manque de répétitions. Cela se traduit par le fait que chaque segment de nos corpus est constitué en moyenne de 82% de mots n'apparaissant qu'une seule fois. Uniquement 11% de mots sont prononcés deux fois dans le même thème. Très peu de mots ont plus de trois occurrences dans le même segment thématique. Le manque de répétition tient essentiellement à la nature de ce genre d'émissions. En effet, une grande partie des journaux d'information est constituée de parole préparée. Les locuteurs professionnels (présentateur principal, reporter, *etc.*) tentent de parfaire leurs interventions : il est souhaitable que leurs discours ne contiennent pas de fautes et peu de répétition de mots. Pour cela, les journalistes prennent des notes voire rédigent entièrement le contenu de leurs interventions. L'enrichissement de l'émission par d'autres informations utiles à la tâche devient alors une priorité à laquelle nous nous attachons dans les chapitres 5 et 6.

2.4 Métriques d'évaluation

Même si les travaux autour de la segmentation thématique sont très nombreux, les métriques d'évaluation utilisées restent presque les mêmes en ignorant leurs défauts. Dans cette partie, nous présentons tout d'abord les trois métriques standards (p_k, *WindowDiff* et *Rappel/Précision*). Nous décrivons ensuite deux métriques d'évaluation que nous avons proposées : $CouvD$ et $CouvN$.

2.4.1 Beeferman p_k

(Beeferman *et al.*, 1997) ont été les premiers à proposer la mesure p_k. Celle-ci repose sur le principe d'une fenêtre glissante de taille k parcourant en parallèle la segmentation de référence et la segmentation automatique. Le paramètre k correspond au nombre d'unités de base à prendre en compte. Dans le cas d'une transcription automatique, l'unité est le de groupe de souffle (GS). Elle se réfère au nombre de phrases ou de paragraphes s'il s'agit de documents textuels.

Soit N le nombre d'unités de base dans l'émission,
r_i et h_i correspondent respectivement à la $i^{ème}$ unité de la segmentation de référence et de l'hypothèse.

La mesure p_k compte le nombre de fois où les deux extrémités de la fenêtre glissante appartiennent au(x) même(s) segment(s), à la fois dans la segmentation de référence et dans celle de l'hypothèse. Elle est donnée par l'expression suivante :

$$p_k(R,H) = \frac{1}{N-k}\sum_{i=1}^{N-k} f(f(r_i, r_{i+k}), f(h_i, h_{i+k})). \tag{2.1}$$

La fonction f est égale à 1 si ses deux arguments sont égaux. Sinon, elle est égale à 0 (opérateur *XNOR*). Pour réécrire la métrique p_k en terme de taux d'erreur notée par p_k^e, il suffit de remplacer, dans l'équation 2.1, la fonction *XNOR* par *XOR*. Ce qui revient à calculer la valeur complémentaire à 1 de p_k (*i.e.* $p_k^e = 1 - p_k$).

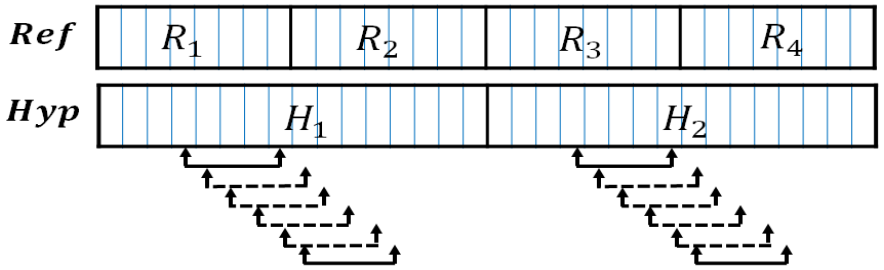


FIGURE 2.6 – *Le principe de la mesure p_k, avec une fenêtre de taille $k = 4$. Les rectangles représentent les unités de base, les lignes pointillées correspondent à une pénalisation et les lignes pleines indiquent qu'aucune pénalité n'est affectée.*

Exemple

La figure 2.6 présente deux segmentations thématiques d'un document de taille 32. La première est produite manuellement et la deuxième est générée automatiquement. En comparant les deux segmentations, nous constatons que le système a commis deux erreurs de type suppression. En effet, deux changements thématiques existent dans la référence qui n'ont pas été détectées par le

système de segmentation. Avec une taille de fenêtre $k = 4$, la mesure p_k enregistre 8 pénalisations (4 pour chaque erreur) ce qui donne $p_k^e(R, H) = \frac{8}{32-4} = 0,28$.

(Pevzner et Hearst, 2002) ont montré que la mesure p_k présente des failles. Parmi lesquelles, on notera :
- **Certaines erreurs ne sont pas pénalisées** : soient n et m le nombre de frontières dans une fenêtre de taille k pour la segmentation de référence et automatique respectivement. Si $n = 1$ et $m = 2$, un changement de thème est détecté alors qu'il n'existe pas. Le système de segmentation commet donc une erreur de type fausse alarme. Cependant, la métrique p_k n'affecte aucune pénalité dans cette situation (voir la figure 2.7). De façon similaire, lorsque $n = 2$ et $m = 1$, le système de segmentation thématique commet une erreur de type suppression. Néanmoins, la fenêtre glissante de taille k ne permet pas de détecter l'erreur. La mesure p_k ne donne aucune pénalisation au système de segmentation pour ce type d'erreurs.

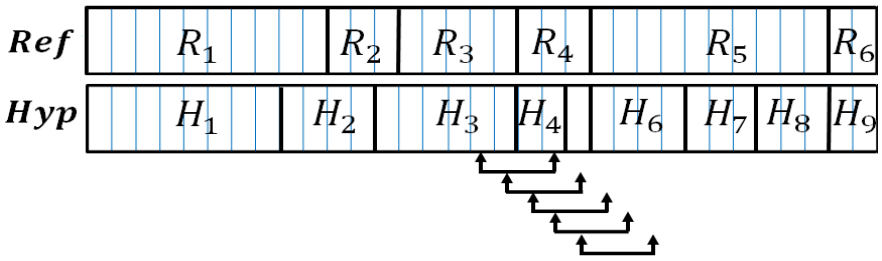


FIGURE 2.7 – *Fausse alarme non pénalisée par la mesure p_k.*

- p_k **est sensible à la taille des segments thématiques** : cette tendance s'exprime particulièrement pour les segments ayant une longueur inférieure à la taille k de la fenêtre. Supposons que A et B sont deux segments adjacents et que le système n'a pas détecté de changement de thème dans cette zone. La métrique p_k désavantage la qualité de la segmentation k fois si $Taille(A) + Taille(B) > 2k$ et la pénalisation diminue linéairement avec la valeur de k si $k < Taille(A) + Taille(B) < 2k$. Ce phénomène est illustré en figure 2.8, en comparant le comportement de p_k avec les deux segmentations H_1 et H_2. Même si les deux segmentations possèdent le même type d'erreur (fausse alarme), elles sont pénalisées différemment. En effet, le score p_k^e de H_2 est le double de H_1. Ceci s'explique par le fait que la mesure p_k pénalise moins les fausses frontières situées à côté d'une frontière de référence.

D'autres problèmes ont été évoqués comme le fait que l'absence ou un décalage d'une frontière provoquent la même pénalité (voir la segmentation H_0 et H_3 en figure 2.8). Par ailleurs, les valeurs de p_k n'ont pas une signification claire.

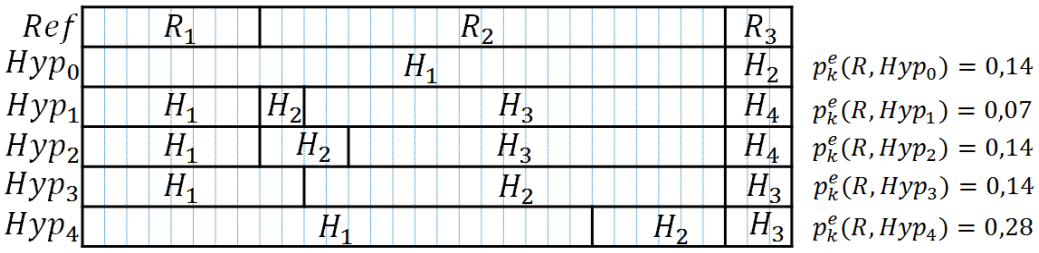


FIGURE 2.8 – *Comportement de p_k vis-à-vis de la taille des segments.*

2.4.2 WindowDiff

Pour contourner certains défauts de la mesure p_k, (Pevzner et Hearst, 2002) proposent la mesure *WindowDiff* (*WD*) qui s'appuie toujours sur le principe de la fenêtre glissante. Si p_k compte le nombre de fois où deux unités de base d'une distance k appartiennent au même segment à la fois dans la référence et dans l'hypothèse, la mesure *WD* calcule la différence du nombre de frontières dans une fenêtre de taille k.

Pour un document de taille N unités, la mesure *WD* est donnée par :

$$WindowDiff(R, H) = \frac{1}{N-k} \sum_{i=1}^{N-k} (\mid b(r_i, r_{i+k}) - b(h_i, h_{i+k}) \mid) \qquad (2.2)$$

où $b(i, j)$ est le nombre de frontières entre les unités i et j dans le document de N unités.

Même si (Pevzner et Hearst, 2002) ont montré que la mesure WD pallie les limites de p_k comme la non ignorance de l'ajout des petits segments, la mesure est loin d'être parfaite. En effet, (Lamprier *et al.*, 2007) signalent que les frontières d'hypothèses situées dans l'intervalle $[1, k]$ (début de l'émission) ou à la fin ($[N-k+1, N]$) sont moins pénalisées par rapport aux autres segments (voir la figure 2.9). Deux solutions ont été proposées par (Lamprier *et al.*, 2007). La première consiste à ajouter k unités fictives au début et à la fin du document. Dans la deuxième solution, les auteurs ont proposé une adaptation de la formule WD pour corriger quelques limites.

(Georgescul *et al.*, 2006a) reprochent à la métrique le fait que les fausses alarmes et le manque de frontières sont pénalisés de la même façon. Ce fonctionnement favorise les systèmes produisant moins de frontières. La fenêtre glissante qui parcourt la segmentation de référence détecte peu de frontières thématiques. Dans le cas où la segmentation automatique produit moins de limites hypothétiques, la mesure *WD* comptera alors peu d'erreurs. Les mêmes auteurs soulèvent le problème de l'interprétation des résultats. Les scores peuvent

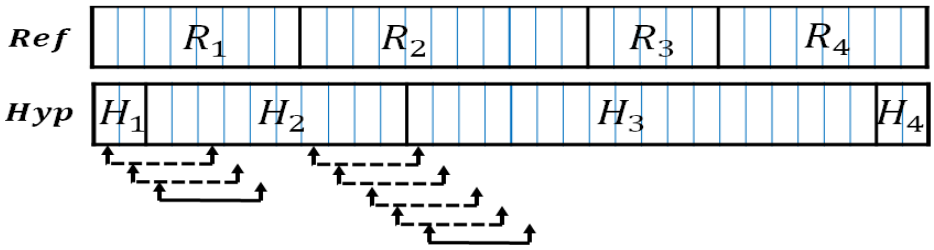


FIGURE 2.9 – *Comportement de WD avec des fausses alarmes situées au début et à la fin.*

prendre des valeurs supérieures à 1 (Sitbon et Bellot, 2004) et ne permettent donc pas de comparer efficacement les résultats.

(Scaiano et Inkpen, 2012) proposent la métrique *WinPR* pour corriger quelques failles de *WD* en s'intéressant beaucoup plus à la position des frontières. De façon similaire à la mesure proposée par (Lamprier *et al.*, 2007), *WinPR* ajoute k unités fictives au début et à la fin du document à évaluer.
Pour chaque fenêtre de taille k, la mesure *WinPR* calcule les 4 cas possibles (présentés ci-dessous). Ces cas peuvent se rencontrer lors de la comparaison du nombre de frontières des deux segmentations :

— Vrai positif (VP) : la segmentation de référence et la segmentation automatique placent une frontière exactement dans la même position.

— Faux positif (FP) : le système annonce un changement de thème alors qu'il n'existe pas dans la référence.

— Vrai négatif (VN) : la segmentation de référence et la segmentation automatique ne décèlent pas de changement de thème.

— Faux négatif (FN) : un changement de thème n'a pas été détecté par le système.

Plus formellement, les quatre cas possibles de la métrique *WinPR* sont données par :

$$VP = \sum_{i=1-k}^{N} \min(r_{i,i+k}, h_{i,i+k}) ; \qquad VN = -k(k-1) \sum_{i=1-k}^{N} (k - max(r_{i,i+k}, h_{i,i+k}))$$

$$FP = \sum_{i=1-k}^{N} \max(0, h_{i,i+k} - r_{i,i+k}) ; \qquad FN = \sum_{i=1-k}^{N} \max(0, r_{i,i+k} - h_{i,i+k})$$

où $r_{i,i+k}$ et $h_{i,i+k}$ sont respectivement les nombres de frontières dans la segmentation de référence et de l'hypothèse d'une fenêtre allant de i à $i+k$.
À partir de ces 4 cas, les mesures *WinR* et *WinP* sont définies. Elles correspondent respectivement aux mesures *Rappel* et *Précision*.
Récemment d'autres métriques ont été également proposées comme avec les travaux (Kazantseva et Szpakowicz, 2012) et (Fournier, 2013).

2.4.3 Mesure Rappel/Précision

La qualité de la segmentation peut être aussi évaluée en comparant la position des ruptures thématiques de référence avec celles à évaluer à l'aide des mesures *Rappel, Précision et Fmesure*. Elles sont définies de la façon suivante :

$$Rappel = \frac{nombre\ de\ frontières\ correctes\ proposées}{nombre\ total\ de\ frontières\ à\ trouver} \tag{2.3}$$

$$Précision = \frac{nombre\ de\ frontières\ correctes\ proposées}{nombre\ total\ de\ frontières\ proposées} \tag{2.4}$$

$$Fmesure = \frac{2 \times Rappel \times Précision}{Rappel + Précision} \tag{2.5}$$

Un rappel de 100% indique que toutes les frontières de référence ont été correctement trouvées. Une précision de 100% signifie que l'ensemble des frontières proposées par le système sont correctes.
L'évaluation du système par les mesures *Rappel/Précision* nécessite de définir d'abord ce qu'est une frontière correcte. Logiquement, une limite thématique proposée par le système est considérée comme correcte si elle coïncide exactement avec un vrai changement de thème. Cette définition n'est pas adaptée à notre tâche car un petit décalage ou l'absence d'une frontière sont considérés comme faux. Or, pour une évaluation juste, une tolérance entre les frontières d'hypothèse et de référence est recommandée. Dans la littérature de la segmentation automatique de documents audio, la tolérance prend une valeur allant de 5 secondes (Rosenberg *et al.*, 2007) à 15 secondes (Xie *et al.*, 2010b). De façon similaire à (Guinaudeau *et al.*, 2010), nous utilisons une tolérance de 10 secondes qui nous semble plus raisonnable d'un point de vue applicatif. L'exemple donné dans la figure 2.10 montre que la tolérance permet de consi-

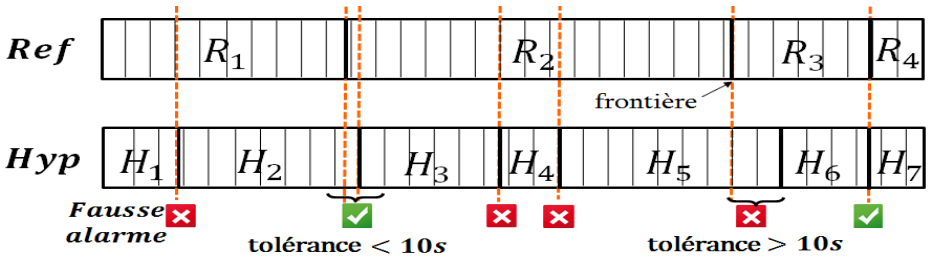


FIGURE 2.10 – *Prise en compte d'un intervalle de tolérance dans le calcul de la mesure Précision/Rappel pour la segmentation thématiques de documents audio.*

dérer la deuxième frontière d'hypothèse comme correcte. Par conséquent, la métrique compte deux frontières correctes parmi les six, ce qui donne un *Rappel* de 66.7

2.4.4 Métriques d'évaluation : *CouvN* et *CouvD*

Les métriques présentées précédemment évaluent la qualité de la segmentation en comparant le positionnement des frontières thématiques de référence avec celles à évaluer soit à l'aide d'une fenêtre glissante (p_k et $windowDiff$) soit par comparaison directe (*Rappel/Précision*). Notons que les scores donnés par la métrique *Rappel/Précision* sont bien adaptés à une application de navigation. En effet, cette métrique renseigne à la fois sur le taux de frontières correctement proposées ainsi que sur le nombre de fausses alarmes. Par exemple, un rappel de 100% et une précision de 40% signifie que l'utilisateur trouvera forcément le début de chaque thème de l'émission. En revanche, 60% des frontières que le système lui proposera ne correspondront pas à des changements de thèmes.

Nous proposons deux métriques exprimant la qualité d'une segmentation thématique, soit par le *nombre* soit par la *durée* des *segments correctement trouvés* :

— *En terme de nombre de segments correctement trouvés* : *CouvN* s'appuie sur le nombre de segments correctement proposé par le système de segmentation. La meilleure segmentation est alors celle qui propose le maximum de segments corrects.

— *En terme de durée des segments correctement trouvés* : *CouvD* s'appuie sur la durée des segments correctement proposés par le système de segmentation. La meilleure segmentation est celle qui propose des segments corrects couvrant la plus grande partie de l'émission.

Ces deux mesures sont complémentaires. En effet, *CouvN* s'intéresse au nombre de segments corrects proposés par le système de segmentation thématique tandis que *CouvD* s'intéresse à la durée totale de segments corrects. Ainsi, *CouvD* ne récompense pas de la même façon un segment correct qui dure une dizaine de minutes d'un autre qui dure quelques secondes. Par ailleurs, les mesures *CouvN* et *CouvD* sont adaptées à plusieurs applications comme la recherche d'information ou le résumé automatique. En effet, ces cas d'usage s'intéressent beaucoup plus aux segments proposés qu'aux instants de changement de thèmes. L'évaluation de l'émission entière convient parfaitement aux applications n'exploitant pas tous les segments d'un document. Par ailleurs, l'évaluation de chaque segment pris séparément intéresse les systèmes n'exploitant qu'une partie de l'émission (comme les moteurs de recherche).

La performance d'une segmentation thématique exprimée selon les trois métriques d'évaluation *Rappel/Précision*, *CouvN* et *CouvD* donne une grande lisibilité sur la performance des systèmes. D'une part, la mesure *Rappel/Précision* permet d'évaluer le système en terme de détection de frontières. D'autres part, *CouvN* et *CouvD* donnent plus de clarté sur la qualité des segments en ajoutant des informations sur le nombre et la durée de ceux-ci. Dans la suite de ce travail,

nous évaluons les performances de notre système avec ces trois mesures.

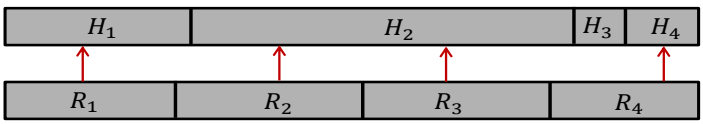
L'évaluation au niveau des segments (*i.e* les mesures *CouvN* et *CouvD*) soulève de nombreuses questions : qu'est-ce qu'un segment correct ? Comment faire la correspondance entre les segments de référence et ceux de l'hypothèse ?

2.4.4.1 Calcul de la couverture entre deux segmentations

La couverture entre deux segments donnés R_i et H_j, notée $Couv_{R_i \leftrightarrow H_j}$ est définie comme la moyenne harmonique de $Couv_{R_i \rightarrow H_j}$ et $Couv_{H_j \rightarrow R_i}$.
$Couv_{R_i \rightarrow H_j}$ est le taux de recouvrement du segment de référence par le segment hypothèse et $Couv_{H_j \rightarrow R_i}$ est le taux de recouvrement du segment hypothèse par le segment de référence. Plus formellement, la couverture entre les deux segments R_i et H_j est donnée par :

$$Couv_{R_i \leftrightarrow H_j} = \frac{2 \times Couv_{R_i \rightarrow H_j} \times Couv_{H_j \rightarrow R_i}}{Couv_{R_i \rightarrow H_j} + Couv_{H_j \rightarrow R_i}}. \tag{2.6}$$

De façon similaire à la mesure *Rappel/Précision*, une tolérance doit être appliquée pour autoriser de petits décalages. Nous considérons qu'un segment H_j est correct si $Couv_{R_i \leftrightarrow H_j}$ est supérieure à un seuil de recouvrement γ. Par exemple, pour la segmentation automatique H, comme illustré dans la figure 2.11, avec un seuil de recouvrement $\gamma = 85\%$, un seul segment (H_1) est considéré comme correct.



R_1 *correspond à* H_1 $Couv_{R_1 \rightarrow H_1} = 100\%$ $Couv_{H_1 \rightarrow R_1} = 86\%$ $Couv_{R_1 \leftrightarrow H_1} = 92\%$
R_2 *correspond à* H_2 $Couv_{R_2 \rightarrow H_2} = 94\%$ $Couv_{H_2 \rightarrow R_2} = 48\%$ $Couv_{R_2 \leftrightarrow H_2} = 63\%$
R_3 *correspond à* H_2 $Couv_{R_3 \rightarrow H_2} = 100\%$ $Couv_{H_2 \rightarrow R_3} = 45\%$ $Couv_{R_2 \leftrightarrow H_3} = 62\%$
R_4 *correspond à* H_4 $Couv_{R_4 \rightarrow H_4} = 38\%$ $Couv_{H_4 \rightarrow R_4} = 100\%$ $Couv_{R_4 \leftrightarrow H_4} = 55\%$

FIGURE 2.11 – *Exemple d'évaluation de la segmentation thématique par nombre de segments corrects.*

2.4.4.2 Évaluation par nombre de segments corrects

Cette évaluation consiste à compter le nombre de segments de référence couvrant suffisamment les segments de l'hypothèse (*i.e.* la couverture est supérieure au seuil imposé). Nous calculons les rapports entre :

— Le nombre de segments jugés corrects sur le nombre total de segments de référence (*Rappel*) noté $\mathcal{R}_\mathcal{N}$.

— Le nombre de segments jugés corrects sur le nombre de segments proposés par le système (*Précision*) noté $\mathcal{P}_\mathcal{N}$.

Les valeurs de $\mathcal{R}_\mathcal{N}$ et $\mathcal{P}_\mathcal{N}$ sont données par :

$$\mathcal{R}_\mathcal{N} = \frac{1}{N_R} \sum_{i=1}^{N_R} \varphi(R_i) \tag{2.7}$$

$$\mathcal{P}_\mathcal{N} = \frac{1}{N_H} \sum_{j=1}^{N_H} \varphi(H_j) \tag{2.8}$$

où : N_R est le nombre de segments de référence,

N_H est le nombre de segments proposés par le système,

$$\varphi(R_i) = \begin{cases} 1 & si \; \exists \; H_j \; telle \; que \; Couv_{R_i \leftrightarrow \; H_j} > \gamma \\ 0 & sinon \end{cases} \tag{2.9}$$

$$\varphi(H_j) = \begin{cases} 1 & si \; \exists \; R_i \; telle \; que \; Couv_{R_i \leftrightarrow \; H_j} > \gamma \\ 0 & sinon \end{cases} \tag{2.10}$$

Ces deux mesures sont synthétisées par le calcul de la *Fmesure* associée. La mesure $CouvN$ est donnée dans l'équation 2.11.

$$CouvN = \frac{2 \times \mathcal{R}_\mathcal{N} \times \mathcal{P}_\mathcal{N}}{\mathcal{R}_\mathcal{N} + \mathcal{P}_\mathcal{N}} \tag{2.11}$$

Remarque : le calcul de couverture est symétrique (*i.e.* $Couv_{R_i \leftrightarrow H_j} = Couv_{H_j \leftrightarrow R_i}$). En d'autres termes, la correspondance entre les segments de l'hypothèse et de référence est la même dans les deux cas : soit on cherche pour chaque segment de référence sa correspondance dans la segmentation d'hypothèse ($\underset{R_i \in R}{\text{argmax}} \, Couv_{R_i \leftrightarrow H_j}$) soit à l'inverse, pour chaque segment de l'hypothèse, on lui attribue le segment de référence qui maximise la couverture ($\underset{H_j \in H}{\text{argmax}} \, Couv_{H_j \leftrightarrow R_i}$).

2.4.4.3 Évaluation par durée de segments corrects

Dans la sous-section précédente, nous avons vu que les fonctions $\varphi(R_i)$ et $\varphi(H_j)$ prennent deux valeurs possibles : 1 si $Couv_{R_i \leftrightarrow H_j} > \gamma$ et 0 sinon. Dans l'évaluation, en terme de durée des segments corrects (*CouvD*), la récompense dépend de la taille des segments, c'est-à-dire que les mesures $\mathcal{R}_\mathcal{N}$ et $\mathcal{P}_\mathcal{N}$ décrites précédemment sont pondérées par la durée des segments relativement à la durée de l'émission. Ce qui donne les formules :

$$\mathcal{R}_\mathcal{D} = \frac{1}{d} \sum_{i=1}^{N_R} d(R_i)\varphi(R_i) \tag{2.12}$$

$$\mathcal{P}_\mathcal{D} = \frac{1}{d} \sum_{j=1}^{N_H} d(H_j)\varphi(H_j) \tag{2.13}$$

où $d(S)$ est la durée du segment S et d la durée de l'émission.

2.4.4.4 Exemple d'évaluation

La figure 2.12 représente la segmentation de référence et de l'hypothèse du journal *TF1_20Heures* du 14/02/2014. Le détail des instants *début* et *fin* de chaque segment est donné dans le tableau 2.5.

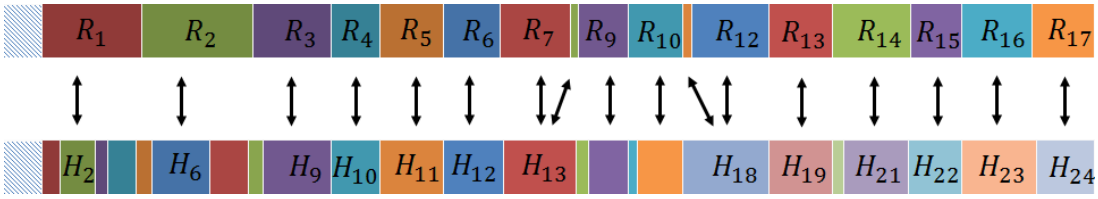


FIGURE 2.12 – *Segmentation de référence et de l'hypothèse d'une émission de* 1915.8 *secondes, la durée totale des segments corrects est de* 1498.4 *secondes.*

La système de segmentation retourne 24 segments, alors que le journal d'information traite 17 thèmes. Chaque segment de référence est mis en correspondance avec le segment de l'hypothèse pour lequel la couverture est maximale. Comme illustré en figure 2.12 par des flèches. Ensuite, les couvertures bi-directionelles sont calculées :

$Couv_{R_1 \leftrightarrow H_2} = 52.0$ $Couv_{R_2 \leftrightarrow H_6} = 68.6$ $Couv_{R_3 \leftrightarrow H_9} = 92.7$
$Couv_{R_4 \leftrightarrow H_{10}} = 99.8$ $Couv_{R_5 \leftrightarrow H_{11}} = 99.8$ $Couv_{R_6 \leftrightarrow H_{12}} = 97.4$
$Couv_{R_7 \leftrightarrow H_{13}} = 93.6$ $Couv_{R_8 \leftrightarrow H_{13}} = 15.2$ $Couv_{R_9 \leftrightarrow H_{15}} = 88.7$

Segmentation de référence			Segmentation de l'hypothèse		
	début	fin		début	fin
R_1	80.86	263.21	H_1	80.86	115.14
R_2	263.21	465.29	H_2	115.14	179.22
R_3	465.29	608.65	H_3	179.22	200.93
R_4	608.65	696.95	H_4	200.93	253.97
R_5	696.95	813.25	H_5	253.97	282.02
R_6	813.25	915.91	H_6	282.02	387.47
R_7	915.91	1043.65	H_7	387.47	458.27
R_8	1043.65	1058.98	H_8	458.27	484.75
R_9	1058.98	1149.21	H_9	484.75	608.56
R_{10}	1149.21	1247.20	H_{10}	608.56	696.79
R_{11}	1247.20	1264.93	H_{11}	696.79	813.14
R_{12}	1264.93	1405.07	H_{12}	813.14	921.37
R_{13}	1405.07	1521.08	H_{13}	921.37	1054.95
R_{14}	1521.08	1663.49	H_{14}	1054.95	1077.25
R_{15}	1663.49	1756.76	H_{15}	1077.25	1149.14
R_{16}	1756.76	1883.84	H_{16}	1149.14	1165.41
R_{17}	1756.76	1996.73	H_{17}	1165.41	1247.18
			H_{18}	1247.18	1404.99
			H_{19}	1404.99	1520.9
			H_{20}	1520.9	1541.79
			H_{21}	1541.79	1659.33
			H_{22}	1659.33	1756.54
			H_{23}	1756.54	1892.07
			H_{24}	1892.07	1996.73

Tableau 2.5 – *Les instants début et fin pour chaque segment de référence et de l'hypothèse.*

$Couv_{R_{10}\leftrightarrow H_{17}} = 91.0$ $Couv_{R_{11}\leftrightarrow H_{18}} = 20.2$ $Couv_{R_{12}\leftrightarrow H_{18}} = 94.0$
$Couv_{R_{13}\leftrightarrow H_{19}} = 99.9$ $Couv_{R_{14}\leftrightarrow H_{21}} = 90.4$ $Couv_{R_{15}\leftrightarrow H_{22}} = 97.7$
$Couv_{R_{16}\leftrightarrow H_{23}} = 93.8$ $Couv_{R_{17}\leftrightarrow H_{24}} = 96.2$

Avec une couverture minimale de 85%, 13 segments sont correctement déterminés. En terme de frontières : sur les 23 limites de l'hypothèses, 14 sont jugées correctes, sachant que l'émission compte 16 frontières thématiques.

Évaluation par nombre de segments corrects :

$$\mathcal{R}_{\mathcal{N}} = \frac{13}{17} = 0.76, \quad \mathcal{P}_{\mathcal{N}} = \frac{13}{24} = 0.54, \quad CouvN = 0.63$$

Évaluation par durée de segments corrects :

$$\mathcal{R}_{\mathcal{D}} = 0.78, \quad \mathcal{P}_{\mathcal{D}} = 0.75, \quad CouvD = 0.77$$

Évaluation au niveau des frontières :

$$R = \frac{14}{16} = 0.87, \quad P = \frac{14}{23} = 0.60, \; Fmesure = 0.71$$

2.4.5 Analyse du comportement de la métrique *CouvN*

Dans cette partie, nous étudions le comportement de la mesure $CouvN$ selon le type d'erreurs (insertion ou suppression) et la taille du segment.

2.4.5.1 Insertion de faux segments

Soit R_i le $i^{ème}$ segment de référence. On suppose que le système de segmentation thématique détecte deux segments H_j et H_k alors que la référence n'en comporte qu'un seul. La métrique $CouvN$ ne pénalise pas de la même façon les fausses alarmes. En effet, si $Couv_{R_i \leftrightarrow H_j} < \gamma$ et $Couv_{R_i \leftrightarrow H_k} < \gamma$, la métrique d'évaluation considère que les deux segments H_j et H_k sont faux. Cependant, si l'un d'eux est correct (la couverture dépasse le seuil γ), un seul segment est enregistré comme erroné. Ce constat est illustré en comparant les quatre segmentations de la figure 2.13. Chacune de ces segmentations d'hypothèse comporte une seule erreur de type insertion.
Pour la segmentation Hyp_1 ($Couv_{R_1 \leftrightarrow H_1} > 85\%$)

$$\mathcal{R}_{\mathcal{N}} = \frac{8}{8}; \quad \mathcal{P}_{\mathcal{N}} = \frac{8}{9}; \quad CouvN = 94.1\%;$$

Pour la segmentation Hyp_2 ($Couv_{R_1 \leftrightarrow H_1} < 85\%$)

$$\mathcal{R}_{\mathcal{N}} = \frac{7}{8}; \quad \mathcal{P}_{\mathcal{N}} = \frac{7}{9}; \quad CouvN = 82.5\%;$$

Pour la segmentation Hyp_3 ($Couv_{R_4 \leftrightarrow H_4} < 85\%$)

$$\mathcal{R}_{\mathcal{N}} = \frac{7}{8}; \quad \mathcal{P}_{\mathcal{N}} = \frac{7}{9}; \quad CouvN = 82.5\%$$

Pour la segmentation Hyp_4 ($Couv_{R_1 \leftrightarrow H_1} < 85\%$, $Couv_{R_1 \leftrightarrow H_2} < 85\%$ et $Couv_{R_1 \leftrightarrow H_3} < 85\%$)

$$\mathcal{R}_{\mathcal{N}} = \frac{6}{8}; \quad \mathcal{P}_{\mathcal{N}} = \frac{6}{9}; \quad CouvN = 70.6\%$$

La métrique $CouvN$ ne pénalise pas de la même façon les segments incorrects. Les scores retournés dépendent de leur emplacement et de leur durée. En effet, avec une seule erreur de type insertion, les segmentations Hyp_2 et Hyp_3 sont pénalisées beaucoup plus que la segmentation Hyp_1. Contrairement au faux segment H_2 retourné dans Hyp_1, les segments H_2 et H_5 respectivement de Hyp_2 et Hyp_3, empêchent leurs voisins d'avoir la couverture minimale demandée. Par conséquent, les deux segments sont considérés comme faux.

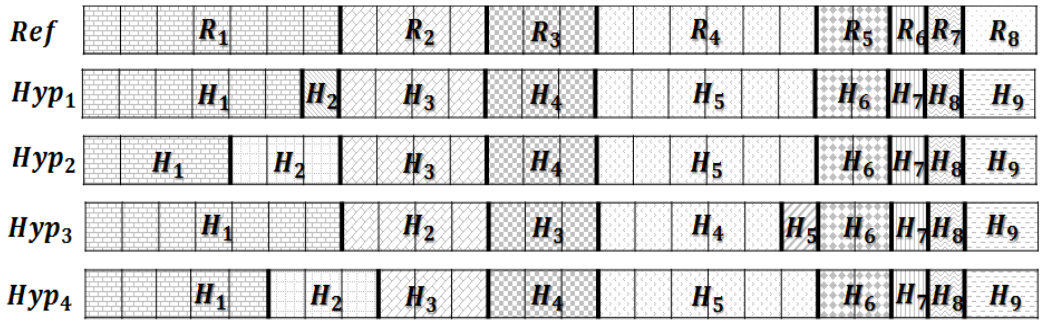


FIGURE 2.13 – *Pénalisation des fausses alarmes.*

Dans les segmentations d'hypothèse Hyp_1, Hyp_2 et Hyp_3, l'insertion de faux segments a été produite dans un seul segment de référence, sans effet de bord. Cependant, le faux segment H_2 de la segmentation Hyp_4 a un impact négatif sur les voisins gauche et droite (H_1 et H_3 respectivement). Par conséquent, la métrique $CouvN$ considère que les trois segments d'hypothèses (H_1, H_2 et H_3) sont erronés.

2.4.5.2 Suppression de segments corrects

Dans certains cas, la mesure $CouvN$ (et par conséquent $CouvD$) est assez souple avec la non détection des segments de petites tailles à condition que le segment de rattachement (son voisin) soit correct.
Par exemple, la métrique d'évaluation compte une seule erreur au niveau de $\mathcal{R}_\mathcal{N}$ pour la segmentation Hyp_5 et considère que tous les segments retournés sont corrects. Avec le même type d'erreur, la pénalité s'élève lorsque la sous-segmentation crée une erreur sur le segment de rattachement. C'est le cas de la segmentation d'hypothèse Hyp_6 de la figure 2.14 , où les segments H_7 et H_8 ont été fusionnés. La métrique d'évaluation juge que la segmentation a commis deux erreurs.
Pour la segmentation Hyp_5 ($Couv_{R_1 \leftrightarrow H_1} > 85\%$)

$$\mathcal{R}_\mathcal{N} = \frac{7}{8}; \quad \mathcal{P}_\mathcal{N} = \frac{7}{7}; \quad CouvD = 93.3\%;$$

Pour la segmentation Hyp_6 ($Couv_{R_9 \leftrightarrow H_8} < 85\%$)

$$\mathcal{R}_\mathcal{N} = \frac{6}{8}; \quad \mathcal{P}_\mathcal{N} = \frac{6}{7}; \quad CouvD = 80.0\%;$$

L'analyse de la métrique $CouvN$, nous a permis de comprendre que

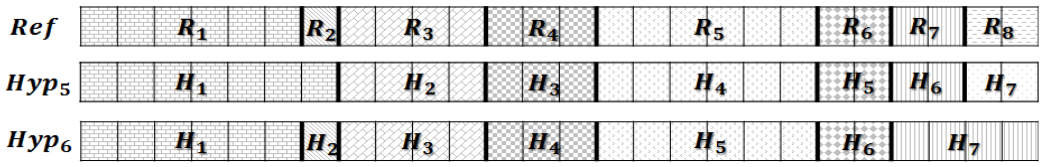


FIGURE 2.14 – *Comportement de CouvN avec la suppression des segments.*

— Les erreurs d'insertion ou de suppression ne passent pas inaperçues (même avec les cas moins graves).

— La sur-segmentation est légèrement plus pénalisée qu'une sous-segmentation si l'erreur n'a aucun impact sur les voisins. Dans le cas inverse, *CouvN* prend alors une valeur un peu plus élevée avec une insertion.

Il faut noter que dans certains domaines comme la recherche d'information et le résumé automatique, il est préférable d'ajouter un segment de petite taille plutôt que de le supprimer.

2.5 Conclusion

Dans ce chapitre, nous avons d'abord présenté les trois corpus sur lesquels sont réalisées toutes les expériences décrites dans ce manuscrit. Les corpus contiennent des journaux d'informations de plusieurs chaînes couvrant divers formats. Ensuite, nous avons introduit les métriques d'évaluation les plus utilisées. Nous avons également mis en lumière leurs points faibles.

Après, nous avons proposé deux mesures complémentaires : *CouvN* et *CouvD*. Elles présentent l'avantage d'avoir des scores facilement interprétables et elles reflètent la performance du système au niveau de chaque segment séparément ainsi qu'au niveau de l'émission globale. Pour les applications directes de la segmentation thématique comme la recherche d'information, le résumé automatique, *etc.* l'évaluation la plus adéquate est celle exprimée en durée de segments. Cependant, pour des fins de navigation, le score exprimé par nombre de segments trouvés est le plus représentatif car l'utilisateur cherche le bon emplacement des frontières. Par ailleurs, la prise en compte conjointe de l'évaluation au niveau des frontières et des segments corrects donne plus de visibilité sur le comportement du système de segmentation thématique. Les deux mesures *CouvN* et *CouvD* permettent de mieux analyser les résultats en étudiant les points forts et faibles du système de segmentation thématique. Ceci permet par exemple de conclure sur les caractéristiques des segments facilement détectables par rapport à la durée des segments ou à la redondance des termes.

Chapitre 3

Système de base et expériences préliminaires

Sommaire

3.1 Introduction

L'état de l'art propose un nombre non négligeable d'algorithmes de segmentation thématique basés sur le calcul de la cohésion lexicale. Même si le degré

de sophistication et d'efficacité varie d'un algorithme à un autre, l'absence d'indices de ruptures thématiques, en particulier la répétition des mots, rend ces algorithmes moins efficaces (notamment ceux qui n'utilisent pas de corpus d'apprentissage). Nous avons choisi TextTiling (Hearst, 1997) comme algorithme de base pour mener nos contributions. Certes, il n'est pas forcément le plus performant, mais il reste le point de départ de plusieurs algorithmes comme (Xie *et al.*, 2008) et (Zheng *et al.*, 2012). De plus, il est non seulement simple à implémenter mais il est aussi facile pour l'interprétation des résultats.

Ce chapitre décrit notre algorithme de base ainsi que les modifications que nous avons apportées à *TextTiling* tel qu'il a été présenté dans (Hearst, 1997). Nous étudions aussi l'influence des paramètres les plus importants de notre algorithme de base. Bien que les contributions majeures que nous avons menées durant cette thèse s'appuient sur l'algorithme décrit dans ce chapitre, nos propositions peuvent être adaptées à d'autres approches.

3.2 Approche retenue : pourquoi TextTiling ?

Dans plusieurs travaux, les techniques de segmentation dédiées au traitement automatique du langage naturel (TAL) ont été appliquées telles quelles aux documents oraux. Rapidement les chercheurs ont observé qu'elles étaient moins efficaces dans ce contexte (Guinaudeau, 2011). Il faut noter que la différence de performance revient essentiellement aux erreurs de transcriptions. En effet, pour accéder au sens des documents oraux, le signal audio de chaque émission est transformé en un texte correspondant aux paroles prononcées. Cette opération est réalisée par le biais d'un système de reconnaissance automatique de la parole. Des erreurs de transcription sont alors susceptibles de perturber le calcul de la cohésion lexicale entre les différentes parties de l'émission.

D'autres difficultés communes aux documents écrits et oraux ont une influence directe sur la qualité de la segmentation. Par exemple, pour les articles de presse et les journaux télévisés, nous pouvons évoquer la spécificité suivante : les auteurs et les journalistes dans ce contexte, essaient d'éviter au maximum la répétition des mots. Pour cela, ils recourent à des synonymes ou à des mots proches sémantiquement. Ce mécanisme a une influence directe sur la performance du système de segmentation thématique. En effet, les algorithmes basés sur la cohésion lexicale considèrent qu'un segment est cohérent s'il est caractérisé par une forte réitération des mots. S'appuyer uniquement sur les mots ayant la même forme écrite limite la fiabilité des valeurs de cohésion, quel que ce soit le système de segmentation thématique.

Dans les travaux menés dans cette thèse, nous nous focalisons davantage sur la représentation vectorielle de l'émission. En effet, la plupart des algorithmes de segmentation basés sur le calcul de la cohésion lexicale exploitent différemment les mots constituant l'émission. Cependant, ces algorithmes peuvent être moins performants s'ils n'utilisent que la répétition des mots pour détecter les changements de thèmes. Nous proposons d'enrichir la représentation vectorielle de l'émission en prenant en compte les points suivants :

1. Plusieurs occurrences d'un même mot peuvent être liées à plusieurs thèmes.
2. Une émission télévisée peut contenir des informations complémentaires à la répétition de mots qui sont utiles pour le calcul de la cohésion.
3. Deux mots peuvent être liés même s'ils ne s'écrivent pas de la même façon.

Comme solution pour le premier point, nous proposons deux approches pour améliorer le processus de pondération des mots (*cf* chapitre 4). Pour le deuxième point, nous proposons d'inclure la distribution de locuteur dans le calcul de la cohésion (*cf* chapitre 5). Comme solution pour le troisième phénomène, nous proposons d'intégrer les relations sémantiques entre les mots dans le calcul de la cohésion lexicale (*cf* chapitre 6).

Il faut aussi noter que c'est la simplicité de *TextTiling* qui nous a encouragé à utiliser cet algorithme pour explorer nos contributions. Ce que nous proposons est aussi applicable sur d'autres algorithmes de segmentation comme *C99, MinCut, etc.*

L'approche de segmentation retenue repose sur la notion de la cohésion lexicale. Notre algorithme comprend quatre phases essentielles et chacune d'elles est constituée d'une ou plusieurs étapes de traitement.

— La première phase consiste à représenter le document audio par des vecteurs de termes pondérés.

— La deuxième phase est celle du calcul de la valeur de similarité qui donne une courbe de cohésion lexicale.

— La troisième phase consiste à sélectionner les instants de changements de thèmes à partir de la courbe générée dans l'étape 2.

— La quatrième étape consiste à effectuer une validation des frontières obtenues précédemment.

À noter que l'algorithme *TextTiling* (Hearst, 1997) ne pondère pas les mots du document. Ainsi, il n'est pas concerné par la phase de validation.

3.3 Représentation vectorielle de l'émission

3.3.1 Pré-traitements

Nous avons appliqué les pré-traitements classiques (voir la section 1.3.1) sur la transcription automatique de l'émission. La lemmatisation des mots a été réalisée à l'aide de l'outil *Lia-tagg* [1] qui est adapté à la langue française. De façon similaire à (Guinaudeau, 2011), le filtrage des mots a été effectué selon leur catégorie grammaticale. Nous avons conservé uniquement les verbes (à l'exception des auxiliaires), les noms et les adjectifs. Dans les documents transcrits automatiquement, chaque mot est généralement accompagné par une mesure de confiance. Plus le score est proche de 1, plus il est probable que le mot soit bien reconnu. Afin d'écarter les mots susceptibles d'être mal transcrits, les mots ayant une mesure de confiance inférieure à la valeur 0.5 sont supprimés. Dans (Guinaudeau, 2011), la mesure de confiance est intégrée directement dans le calcul de la cohésion lexicale. Ce mécanisme permet de pénaliser les mots potentiellement mal reconnus. Nous n'avons pas suivi cette approche, un seuil a été cependant appliqué sur les scores de confiance.

3.3.2 Pondération des termes

En recherche d'information, la pondération est une étape incontournable. Elle consiste à associer un poids à chaque terme d'un document donné. On dit qu'un terme t est important dans le document d s'il est fréquent dans d et s'il n'est pas uniformément réparti dans les autres documents du corpus. La pondération *TF-IDF* estime le poids d'un mot à partir de son importance au niveau local (document) et global (collection).
Au niveau local, un mot fréquent considéré comme important se traduit par une valeur élevée de TF. Au niveau global, un mot se trouvant dans la plupart des documents prend une valeur IDF proche de zéro. En revanche, un mot apparaissant rarement dans d'autres documents de la collection peut avoir un poids élevé.
Le poids du terme t dans le document d est donné par :

$$w_{TF-IDF}(t,d) = TF(t,d) \times IDF(t) \tag{3.1}$$

avec $TF(t,d)$ la fréquence du terme t dans le document d,

$$IDF(t) = \log(\frac{N}{n_t}),$$

1. http ://lia.univ-avignon.fr/chercheurs/bechet/download_fred.html

où N est le nombre total de documents,

n_t est le nombre de documents dans lequel le terme t apparaît.
Autour de *TF-IDF* plusieurs extensions ont été proposées comme la pondération *Okapi* (Robertson et Jones, 1976) donnée par la formule suivante :

$$w_{Okapi}(t,d) = \underbrace{\frac{TF(t,d) \times (k+1) + k}{(1 - b + b \times dl(d)/dl_{avg})}}_{TF_{Okapi}(t,d)} \times \underbrace{\log \frac{N - n_t + 0.5}{n_t + 0.5}}_{IDF_{Okapi}(t)} \tag{3.2}$$

où k et b sont des constantes fixées respectivement à 2 et 0.75,

$dl(d)$ est la longueur du document d (nombre de lignes) et dl_{avg} est la longueur moyenne des documents.

Un journal d'information est un *mono-document* traitant un ensemble de thèmes non liés entre eux (indépendants). La pondération consiste à donner plus de poids aux mots importants dans un thème par rapport aux autres thèmes traités dans l'émission.
(Malioutov et Barzilay, 2006) proposent de pondérer les termes du document sans aucune information externe. Les auteurs divisent *a priori* l'émission en N morceaux de taille identique que nous nommons *chunks*. Chaque chunk est composé d'une succession de groupes de souffle. La notion de chunk correspond alors à la notion de document en recherche d'information. Le terme t dans le groupe de souffle x est associé au poids $w(c(x),t)$, avec $c(x)$ le chunk contenant le groupe de souffle x :

$$w(c(x),t) = TF(c(x),t) \times IDF(t) \tag{3.3}$$

avec TF(c(x),t) la fréquence du terme t dans le chunk $c(x)$.

et $IDF(t) = \log(N/n_t)$

n_t est le nombre de chunks dans lequel le terme t apparaît.
Le principe est illustré dans la figure 3.1.

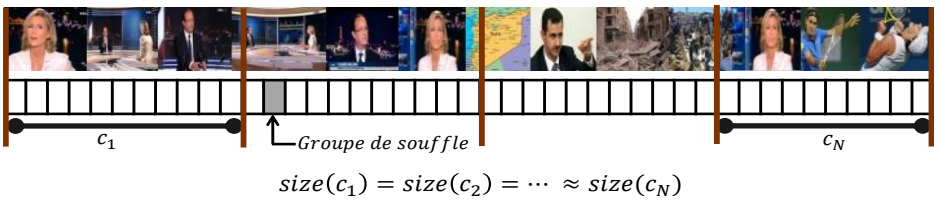


FIGURE 3.1 – *Découpage de l'émission en N chunks uniformes*

Le nombre de chunks N est calculé automatiquement pour chaque émission

en fonction de sa durée et de la durée moyenne des thèmes de l'ensemble des émissions. Cette dernière prend la valeur de 106 secondes qui correspond à la durée moyenne des segments thématiques de nos corpus.

3.3.3 Représentation vectorielle

À l'issue de cette étape, l'émission est transformée en une matrice de taille $n \times m$ où n est le nombre de mots (après les pré-traitements) et m est le nombre de groupes de souffle (voir la figure 3.2). Cette représentation présente certains

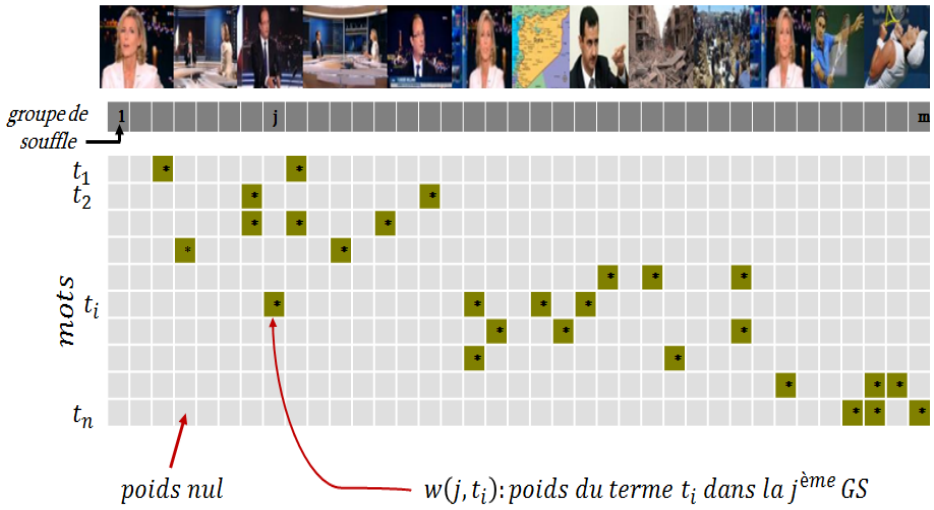


FIGURE 3.2 – *Représentation vectorielle de l'émission.*

avantages :

— Elle est relativement simple à comprendre.

— Elle permet de conserver toutes les informations liées au document.

— Elle supporte les opérations vectorielles simples (produit, somme, *etc.*) et permet entre d'autres d'intégrer les relations sémantiques, d'ajouter d'autres informations, *etc.*

3.4 Calcul de la cohésion lexicale

Comme pour l'algorithme *TextTiling*, la similarité est calculée entre chaque paire de blocs adjacents. Ce calcul s'effectue tout au long de l'émission à l'aide d'une fenêtre glissante de taille $2 \times K$, entre des blocs adjacents de K groupes de souffle de part et d'autre de chaque frontière potentielle. Une valeur élevée

de cohésion signifie que les deux blocs traitent la même thématique. Une faible valeur signifie qu'il y a peu de mots en commun et les deux blocs abordent donc probablement des thèmes différents.
Le poids associé au terme t dans le bloc b est obtenu en sommant les fréquences pondérées du terme t dans chaque groupe de souffle du bloc :

$$v(b,t) = \sum_{x \in b} (f_{x,t} \times w(c(x),t)) \tag{3.4}$$

où $f_{x,t}$ est la fréquence du terme t dans le groupe de souffle x.

Remarques :

— Chaque frontière entre deux groupes de souffle consécutifs est une frontière thématique potentielle.

— Même si les blocs et les chunks sont constitués d'une succession de groupes de souffle, ils sont considérés comme deux notions différentes. Les chunks sont utilisés dans la pondération des termes, chacun d'eux correspond à un document en recherche d'information. Les blocs, eux, sont utilisés pour calculer la cohésion lexicale entre les différentes parties de l'émission.

Afin de calculer la cohésion, nous utilisons la similarité *cosine* qui permet de calculer la similarité entre deux vecteurs de n dimensions en déterminant le cosinus de leur angle. Pour une frontière potentielle j entre deux blocs b_j et b_{j+1}, la similarité est donnée par :

$$cohesion(j) = \frac{\sum_t \left(v(b_j,t) \times v(b_{j+1},t)\right)}{\sqrt{\sum_t \left(v(b_j,t)\right)^2} \times \sqrt{\sum_t \left(v(b_{j+1},t)\right)^2}} \tag{3.5}$$

3.5 Détection des frontières

3.5.1 Recherche de frontières candidates

À partir des valeurs de cohésion, une courbe est tracée pour repérer les instants de changements de thèmes. Pour cela, plusieurs stratégies ont été introduites dans l'état de l'art. Dans (Claveau et Lefèvre, 2011), les auteurs proposent d'appliquer l'algorithme dit « Ligne de Partage des Eaux » (LPE) issu de la morphologie mathématique. Il consiste à simuler l'inondation progressive du relief par ses minima, et à séparer les différents bassins associés à chaque minimum par des digues. Ces dernières, représentent les lignes de partage des eaux, ou autrement dit, les frontières des régions.

Dans l'approche classique de (Hearst, 1997), la profondeur de chaque vallée est calculée en sommant les deux différences obtenues entre la vallée et le pic à gauche puis la vallée et le pic à droite [2] . La profondeur de la vallée a_i notée par $depth(a_i)$ est donnée par :

$$depth(v_i) = (y_{p_{i-1}} - y_{v_i}) + (y_{p_{i+1}} - y_{v_i}) \tag{3.6}$$

où y_{p_c} est la valeur de cohésion du pic ayant la position c.
y_{v_c} est la valeur de cohésion de la vallée ayant la position c.

Nous avons constaté que le lissage tel qu'il est défini dans l'algorithme de TextTiling dégrade les performances du système. En effet, le lissage a pour effet de décaler artificiellement les vallées, voire de les supprimer s'il y a un pic important à proximité. Pour éviter les phénomènes locaux, nous ignorons dans le calcul de la profondeur, les vallées entourées par un pic de petite taille.

Pour améliorer la robustesse de l'extraction des frontières, nous exploitons conjointement la similarité lexicale et la profondeur des vallées. En effet, nous avons observé que la recherche directe sur les valeurs de similarité n'est pas optimale (Bouchekif *et al.*, 2013b). Un changement de thème pour un autre thème proche peut se traduire par une similarité relativement importante. De même, travailler uniquement sur la profondeur des vallées n'est pas optimal : les pics, de part et d'autre d'une vallée, peuvent ne pas être très hauts si un thème contient peu de répétitions de termes. Par exemple, dans la figure 3.3, même

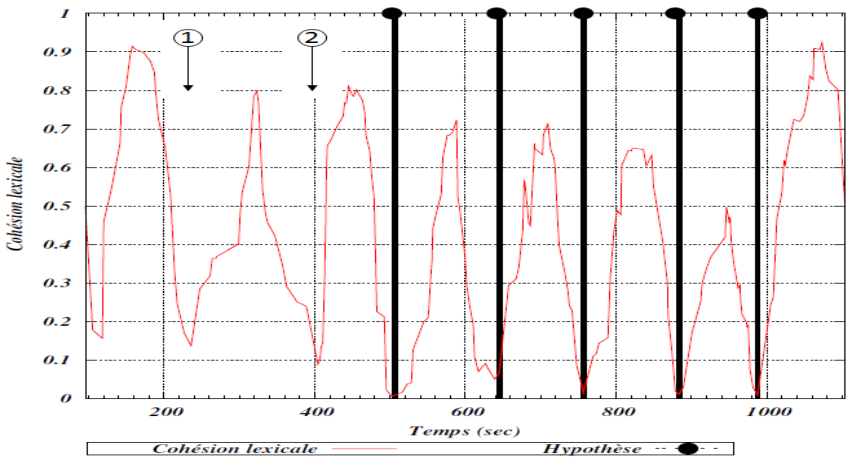


FIGURE 3.3 – *Limites de la détection des frontières basée uniquement sur la profondeur des vallées.*

si les vallées (1) et (2) sont profondes (elles dépassent le seuil fixé), elles ne correspondent pas à un changement de thème. Cependant, les valeurs de cohésion

2. Les définitions des pics et les vallées sont données dans la section 1.4.1.

sont relativement élevées au niveau de ces vallées. Donc la valeur de la cohésion est un indicateur à prendre en compte. Nous proposons ainsi de combiner ces deux mesures complémentaires à l'aide d'une interpolation linéaire. Pour une frontière potentielle j, le score suivant doit être maximisé :

$$score(j) = \lambda(1 - Cohesion(j)) + (1 - \lambda)depth(j) \tag{3.7}$$

Cette combinaison permet de favoriser les valeurs faibles de la cohésion lexicale qui sont en même temps des minima locaux. Par exemple, les vallées (1) et (2) de la figure 3.3 ne seront pas considérées comme des frontières thématiques, selon cette mesure composite.

3.5.2 Sélection parmi les frontières candidates

Au lieu d'appliquer un simple seuillage (comme il a été proposé dans l'algorithme de *TextTiling*), nous proposons un algorithme de division récursive lors duquel est définie une zone d'exclusion autour des frontières trouvées à chaque itération. Le partitionnement consiste à construire un ensemble S de segments. Initialement, S contient un seul segment constitué de l'émission entière.

1. Chaque élément de S est coupé en deux, le point de coupure correspond à la valeur maximale du score dépassant un seuil donné.
2. Les frontières candidates autour du point de coupure (zone d'exclusion) ne seront pas prises en considération lors de la prochaine itération.
3. Les segments obtenus sont présentés à l'étape 1.

L'étape 2 permet de limiter les phénomènes de maxima locaux et garantit que l'on n'obtiendra pas de frontières consécutives. La zone de neutralisation est fixée à 3 groupes de souffle de part et d'autre d'une frontière. L'algorithme s'arrête lorsqu'aucun point de coupure candidat ne dépasse le seuil. Cette approche s'est avérée plus efficace qu'un simple lissage de la courbe pour limiter l'effet des maxima locaux. La granularité des groupes de souffle est trop grande pour envisager un lissage efficace sans perte d'information.

3.6 Validation de la segmentation

Afin d'affiner la segmentation, nous proposons d'ajouter une étape de validation. Elle permet de confirmer ou de rejeter les segments obtenus afin d'assurer que les deux segments situées à gauche et à droite de chaque frontière sont thématiquement différents. À partir d'un ensemble de n segments, nous essayons de les regrouper en m segments (tels que $n \geq m$) thématiquement plus

homogènes que les segments de départ. À la différence de la première étape de l'algorithme, la cohésion est ici calculée entre blocs de tailles différentes. Le principe est le suivant :

1. à partir d'une segmentation, la cohésion est calculée entre tous les segments adjacents.
2. la frontière donnant la valeur la plus élevée de cohésion est supprimée si cette valeur dépasse un certain seuil (*i.e* les segments gauche et droite sont regroupés).
3. l'ensemble des segments d'hypothèse est mis à jour. Les étapes (1) et (2) se répètent jusqu'à ce qu'il n'existe plus de valeur de cohésion dépassant le seuil.

3.7 Évaluation et discussion

3.7.1 Taille de la fenêtre

Comme notre algorithme est basé sur une fenêtre glissante paramétrable, il convient d'étudier son impact sur la performance de notre système de segmentation. La figure 3.4 donne la performance du système en terme de rappel/précision. Pour avoir l'impact voulu (*i.e* uniquement celui de la taille de la fenêtre), nous avons mis à l'écart l'étape de validation. Les courbes sont obtenues en faisant varier la taille de la fenêtre qui prend des valeurs allant de 10 à 26 groupes de souffle par pas de 4.

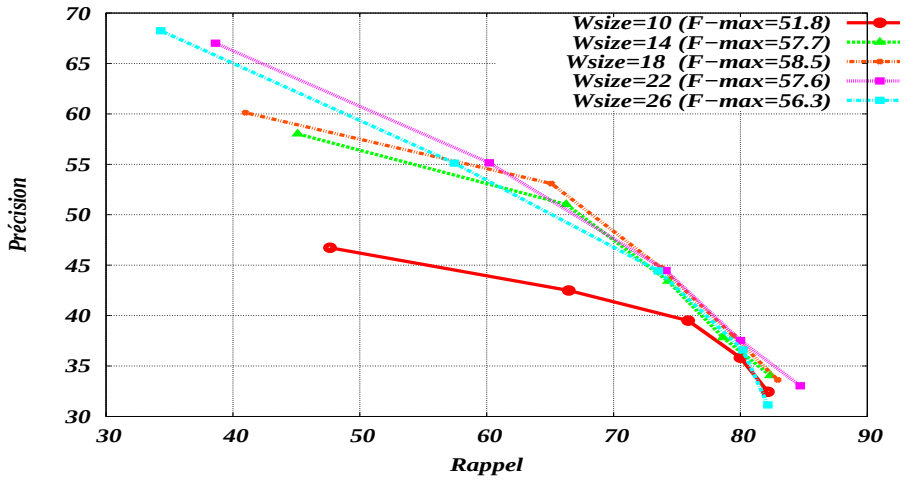


FIGURE 3.4 – *Influence de la taille de la fenêtre sur le corpus MC-0813.*

De l'analyse des courbes, on constate que la taille de la fenêtre ne doit être ni trop petite (de 10 à 14 GS) ni trop grande (de 22 à 26 GS). En effet, une petite fenêtre implique beaucoup de fausses alarmes, surtout si les groupes de souffle contiennent peu de mots et de répétitions. En revanche, une grande fenêtre prendrait en compte des éléments qui ne font pas partie du même segment thématique lors de la segmentation (en particulier les zones contenant une succession de courts segments). Avec une taille de 18 GS, la *F-mesure* prend sa valeur maximale dans le corpus *MC-0813*. Avec une taille moyenne de la fenêtre il y a moins de chance qu'elle couvre plus de deux segments thématiques (*i.e* moins de bruit par rapport à une fenêtre de grande taille). Nous constatons que les valeurs de *rappel* et *précision* sont assez équilibrées pour une fenêtre de taille moyenne. Par la suite, le paramètre *taille de fenêtre* est fixé à 18 groupes de souffle.

3.7.2 Impact du regroupement

Dans cette section, nous présentons les résultats obtenus avec et sans étape de *validation*. Nous testons l'étape de validation en faisant varier le seuil de segmentation, noté s_α. Ceci permet de produire plusieurs hypothèses de segmentation comprenant chacune un nombre différent de frontières.
Les chunks servant à pondérer les termes de l'émission ont été obtenus en fixant le seuil de segmentation. Celui-ci donne un nombre de segments plus proche de ceux de la référence. Le seuil de segmentation optimal estimé sur le corpus de développement est de l'ordre 0.88.
Ces expériences permettent d'observer la capacité de l'étape de validation à détecter les fausses alarmes. Les résultats sont illustrés par la figure 3.5.

En appliquant ce principe, notre algorithme prend en entrée deux ensembles de segments. Le premier sert à la pondération (pour définir la taille des chunks) et le deuxième est utile pour l'étape de regroupement. Les courbes *rappel/précision* ont été obtenues en fixant à chaque fois la valeur de s_α correspondant à la courbe rouge et nous avons fait varier celui du regroupement noté s_β. Ce dernier prend des valeurs allant de 0.04 à 0.24 ainsi que la valeur 1.0 (*i.e* pas de regroupement possible). L'apport de l'étape de validation est remarquable avec les segmentations contenant un nombre important de frontières. En effet, avec un seuil de 0.82, la F-mesure augmente de 47.8% à 58.4%. Cependant même avec un seuil élevé ($s_\alpha = 0.90$), le système arrive à améliorer la F-mesure de 0.4% [3]. Cette différence de gain provient essentiellement du nombre de segments présents

3. Sans validation le système obtient : *R*=41.0, *P*=60.1 et *F-mesure*=48.7, l'intégration de l'étape de validation donne les performances suivantes : $R = 38.8$, $P = 66.9$ et *F-mesure* $= 49.1$

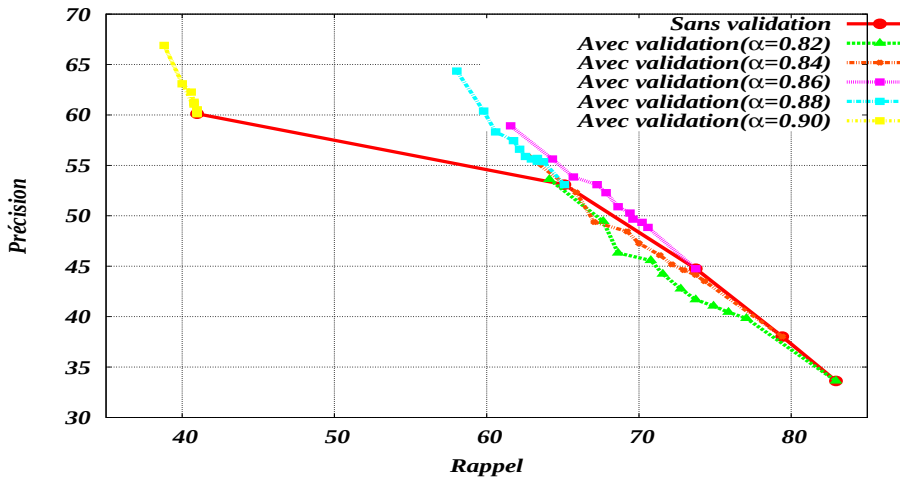


FIGURE 3.5 – *L'apport de l'étape de regroupement sur le corpus MC-0813.*

dans l'étape de validation. Avec des petites valeurs du s_α, le système de segmentation propose un nombre très important de frontières ce qui se traduit par des valeurs de rappel extrêmement fortes mais avec une faible précision. Cela veut dire que l'ensemble des segments présents dans la phase de validation contient la plupart des bonnes frontières. Néanmoins, elles sont entourées par un nombre considérable de fausses alarmes. L'étape de regroupement raffine davantage les hypothèses, c'est pourquoi la précision augmente. Toutefois certaines bonnes frontières sont également supprimées. Plus la valeur du seuil de segmentation augmente, plus le système a la capacité de ne garder que les hypothèses probables, celles qui constituent potentiellement des ruptures. Ainsi, peu de fausses alarmes sont présentes à l'étape de validation. Au total, 781 bonnes frontières ont été supprimées si la valeur s_α vaut 0.82 et seulement 51 frontières avec un seuil de segmentation ayant pour valeur 0.90.

Nous proposons maintenant d'étudier l'influence de l'étape de validation au niveau des segments. Pour cela, nous choisissons les paramètres optimaux estimés sur le corpus de développement (*MC-0813*) : une taille de fenêtre de 18 groupes de souffle et un s_α qui vaut 0.88. Pour le système de segmentation qui intègre la dernière étape, nous appliquons un seuil de regroupement ayant pour valeur 0.08.
Le tableau 3.1 illustre les résultats obtenus avec les trois métriques d'évaluations. L'apport de l'étape de validation est remarquable au niveau du nombre et de la durée des segments corrects (*i.e.* $Couv_N$ et $Couv_D$ respectivement). Au total, 101 segments ont été supprimés dans l'étape de validation, permettant d'augmenter la mesure $\mathcal{P}_\mathcal{N}$, passant de 36.6 à 42.8. Cependant, une légère amélioration a été enregistrée en $\mathcal{R}_\mathcal{N}$. Il faut noter que la dernière étape de l'algorithme supprime parfois des segments corrects.

	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Sans validation	65.1	53.1	58.5	45.4	36.6	40.5	51.8	51.2	51.5
Avec validation	60.6	58.3	59.4	45.6	42.8	44.2	53.8	58.3	58.3

Tableau 3.1 *– Performance du système sans et avec étape de validation dans le corpus MC-0813*

Nous relançons les mêmes expériences précédentes concernant l'étape de validation mais cette fois-ci avec le corpus *MCS7-14*. Les résultats du tableau 3.2 confirment les tendances observées sur le corpus *MC-0813*. C'est-à-dire que l'étape de validation permet de réduire le nombre de fausses alarmes.

	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Sans validation	60.7	52.0	56.0	44.2	36.6	39.6	49.1	49.1	49.1
Avec validation	57.9	59.0	58.4	44.3	44.3	44.3	52.6	53.1	52.8

Tableau 3.2 *– Performance du système sans et avec étape de validation sur le corpus MCS7-14*

3.8 Conclusion

Dans ce chapitre, nous avons présenté notre algorithme de base reposant sur le principe de *TextTiling* que nous avons adapté à la transcription automatique de journaux télévisés. Nous avons notamment apporté des modifications sur le processus d'extraction des frontières thématiques à partir de la courbe de cohésion. L'approche proposée consiste à effectuer une division récursive lors de laquelle est définie une zone d'exclusion autour des frontières trouvées à chaque itération. À partir de la courbe de cohésion lexicale, nous pouvons exploiter soit les valeurs de cohésion, soit les vallées. Dans ce travail, nous proposons de faire une interpolation linéaire des deux.

La pondération uniforme proposée dans l'algorithme de *MinCut* a été adaptée (estimation automatique du nombre de chunks) afin de pénaliser les mots non discriminants. Nous avons aussi décrit le mécanisme de regroupement de segments qui permet d'affiner la segmentation thématique.

Bien évidemment, cette première adaptation de TextTiling fournit des résultats peu satisfaisants pour notre cadre applicatif. Cela est dû à l'utilisation

stricte de répétitions de mots, ce qui pousse le système à proposer des changements thématiques aux mauvais endroits. Nos contributions, présentées dans les chapitres suivants, visent à corriger certaines lacunes de ce système de base.

Deuxième partie

Enrichissement de l'espace vectoriel des documents

Chapitre 4

Pondération intra-document

Sommaire

4.1 Introduction

Les méthodes de segmentation thématique basées sur la distribution des termes considèrent un important changement de vocabulaire comme étant une éventuelle transition thématique. Pour cela, ces méthodes calculent la cohésion lexicale entre différentes sous-parties de l'émission (groupe de souffle, bloc, *etc.*). Idéalement, deux sous-parties traitant deux thèmes différents partagent peu de mots en commun voire pas du tout. Cependant, certains mots apparaissent tout au long de l'émission et peuvent laisser croire qu'il n'y a que peu de frontières alors qu'en réalité ces mots peuvent être communs à plusieurs thèmes. Face à ce problème, deux approches ont été utilisées par les systèmes de segmentation thématique :

— *Par élimination* : les méthodes de cette approche visent à supprimer les mots non discriminants, c'est-à-dire les mots dont le contenu informatif ne permet pas de distinguer un thème par rapport aux autres. Par exemple, la *conjecture de Luhn* (Luhn, 1957) évalue l'importance d'un mot dans un document à partir sa fréquence. Les termes dont la fréquence est en dessous du seuil minimal feq_{min} sont considérés comme trop rares et donc n'ont pas de pouvoir discriminant. Les termes dont la fréquence est au-dessus du seuil maximal feq_{max} sont considérés trop communs. Par conséquent, ils ne peuvent pas discriminer un thème par rapport à l'ensemble des thèmes évoqués dans l'émission. Seuls les mots ayant des fréquences dans l'intervalle $[feq_{min}, feq_{max}]$ sont considérés informatifs sur le contenu du document.

  D'autres systèmes de segmentation thématique utilisent des *stop-listes* pour effectuer un filtrage des mots outils (pour plus de détails voir la sous section 1.3.1).

— *Par pondération* : les méthodes de cette approche donnent un poids à chaque mot de l'émission. Lorsqu'un mot se trouve dans peu de thèmes et en même temps apparaît plusieurs fois au sein d'un même thème, il obtient un score élevé. À l'inverse, il est alors considéré comme non discriminant et est associé à un score faible. Cette définition fait référence aux mesures de pondération standard comme *TF-IDF* et *Okapi*, proposées initialement pour la recherche d'information (voir la sous-section 3.3.2).

Nous notons que certains systèmes de segmentation regroupent les deux approches. L'approche *par élimination* est généralement intégrée dans la phase de pré-traitements pour supprimer les mots outils, tandis que l'approche *par pondération* donne un poids à chaque mot de l'émission.

FIGURE 4.1 – *Distribution des mots dans deux thématiques proches.*

La figure 4.1 illustre un exemple simple de la distribution des mots d'une émission composée de deux thèmes. Le premier porte sur la visite du président Français aux USA et le deuxième traite de la participation de Nicolas Sarkozy au meeting de Nathalie Kosciusko-Morizet.
Dans le premier thème, les mots *président, François, Hollande, français* apparaissent plusieurs fois. Ceci laisserait penser qu'ils ont un lien thématique fort avec le premier thème. Or, ils apparaissent également dans le deuxième thème, ce qui diminue leur pouvoir discriminant. Pour ne pas établir un lien thématique entre les deux segments, les mots figurant dans les deux thèmes devront avoir des poids faibles. Alors que les mots comme *Obama, état, MoDem, Paris, Sarkozy, etc.* apparaissent eux, uniquement dans l'un des deux thèmes. Ils peuvent être alors considérés comme discriminants et devront recevoir un poids fort.

Dans le cas de thèmes très proches, la pondération permet également de mettre en évidence les mots discriminants. Prenons l'exemple illustré dans la figure 4.2, le premier thème traite de la victoire de Syriza. Il considère les mots *droite, partie, sièges, etc.* comme importants. Dans le deuxième segment thématique qui porte sur la dette de la Grèce, les mots qui caractérisent le thème sont *euros, milliards, dette, etc.* Cependant, certains mots sont présents dans les deux thèmes (*syriza, alexis et grèce*) et sont considérés comme moins discriminants.

les électeurs de syriza mettent tous leurs espoirs dans alexis tsipras le nouvel homme premier ministre de la grèce
après le décompte de la moitié des bulletins le parti de gauche radicale syriza l'emporte avec 7,5 points d'avance sur le parti de droite arrivé deuxième
le nouveau alexis tsipras a annoncé la formation d'un gouvernement de coalition avec anel parti de droite souverainiste
sur 53,61% des bulletins dépouillés syriza obtenait 35,54% des voix et 145 sièges sur 300 au parlement contre 28,11% et 75 sièges pour la droite

au cœur du programme de syriza son leader alexis va devoir renégocier la dette de la grèce
concrètement qu'est-ce que ça veut dire la dette de la grèce atteint 300 milliards d'euros près de deux fois la richesse du pays ses créanciers sont les investisseurs privés le fmi et la bce pour 95 milliards d'euros le reste 205 milliards d'euros a été prêté directement ou indirectement par les états de la zone euro
dans le détail l'allemagne c'est 60 milliards d'euros la france 40 milliards d'euros pour chaque français ça représente en moyenne 600 euros de dette supplémentaire

FIGURE 4.2 – *Distribution des mots dans deux segments très proches thématiquement. Le premier porte sur la victoire de Syriza. Le deuxième parle de la dette de la Grèce.*

En règle générale, la pondération permet de donner un poids fort aux mots très fréquents dans un thème et peu fréquents dans le reste de l'émission. Par ailleurs, les mots aussi fréquents dans un thème que dans d'autres auront un poids plutôt faible.

Dans ce chapitre, nous proposons deux variations de la pondération *chunk uniforme* proposée par (Malioutov et Barzilay, 2006) que nous avons décrite dans la sous-section 3.3.2.

4.2 Pondération des termes pour la segmentation thématique

4.2.1 Principe général

Les journaux d'information sont des documents multi-thèmes, c'est-à-dire que plusieurs thèmes sont abordés dans une seule émission. En segmentation thématique, les poids des termes sont estimés soit à partir d'un large corpus, soit à partir du contenu lui-même.

Pondération à partir d'un large corpus
Ce type de pondération nécessite des données textuelles pour faire émerger les mots apparaissant dans plusieurs documents. Par exemple, (Guinaudeau et Hirschberg, 2011) utilisent l'outil *kiwi* (Lecorvé *et al.*, 2008) produisant des poids estimés à partir d'une collection de 800000 articles du journal Le Monde. De façon similaire, (Kern et Granitzer, 2009) pondèrent les mots du document à l'aide d'un corpus et proposent une variante de *TF-IDF*.

Pondération intra-document
Ce type de pondération a été utilisé dans plusieurs travaux. On peut citer les travaux de (Richmond *et al.*, 1997) où le poids $w(t)$ du mot t est donné par la formule suivante :

$$w(t) = \frac{1}{\eta} \times \sum_{i=1}^{n} \arctan(\frac{D_{t,i}}{N / f_t}) \tag{4.1}$$

où $D_{t,i}$ est la distance entre le mot t et sa $i^{ème}$ occurrence la plus proche. Par exemple, $D_{x,2}$ représente la distance entre le mot x et sa deuxième occurrence la plus proche,
N est le nombre total de mots dans le document,
f_t est le nombre d'occurrences du mot t,
η est défini de manière empirique pour que les poids soient entre 0 et 1.

(Dias *et al.*, 2007) proposent une combinaison linéaire de trois scores pour évaluer l'importance de chaque mot dans le document. Le premier score est une variante de *TF-IDF* dont les poids sont donnés par l'équation 4.2 :

$$tf.isf(w) = \frac{stf(w;s)}{|s|} \times \frac{N_s}{sf(w)} \tag{4.2}$$

où $stf(w;s)$ est le nombre d'occurrences du terme t dans la phrase s,
$|s|$ est le nombre de mots dans s,
$sf(w)$ est le nombre de phrases où le terme w apparaît,
N_s est le nombre de phrases dans le document.
Nous pouvons constater que le calcul de $tf.isf$ suit le même principe que celui

de la pondération par chunks [1] uniforme proposée par (Malioutov et Barzilay, 2006). La seule différence est qu'un chunk dans (Dias *et al.*, 2007) correspond à une seule phrase et non pas à un ensemble de phrases obtenues par découpage du document en N chunks de taille uniforme.

Ce choix de limiter la taille de chunk à une unique phrase ne permet pas de prendre en compte le positionnement des mots dans le document. En effet, un mot qui se trouve dans trois phrases consécutives aura le même poids que s'il était placé dans trois phrases qui seraient au début, au milieu et à la fin du document. Pour prendre en considération ce positionnement, (Dias *et al.*, 2007) propose de calculer une *densité* pour chaque mot. L'idée est de donner plus de poids aux mots apparaissant dans des phrases proches et de pénaliser les mots appartenant à des phrases éloignées. La densité du mot w est donnée par :

$$dens(w) = \sum_{k=1}^{|w|-1} \frac{1}{ln(dist(occur(k), occur(k+1)) + e)} \tag{4.3}$$

où $|w|$ est le nombre d'occurrences du mot w,
$occur(k)$ est le positionnement de la $k^{ème}$ occurrence du mot w,
$dist(occur(k), occur(k+1))$ est la distance entre deux occurrences du même mot de positions k et $k+1$.

La pondération finale selon (Dias *et al.*, 2007) est calculée à partir d'une interpolation linéaire entre les scores *TF-IDF(w)*, $tf.isf(w)$ et $dens(w)$, où $tf.idf(w)$ correspond au poids du mot w dans le document par rapport à la collection des documents (la pondération classique en recherche d'information).

4.2.2 Importance de la pondération des termes dans la segmentation thématique

L'étape de pondération n'a pas fait de consensus sur son importance. Par exemple, (Hearst, 1997) et (Huet, 2007) estiment que l'utilisation de la pondération est moins primordiale dans la tâche de segmentation thématique que la fréquence des mots, qu'ils estiment plus robuste. Ils considèrent que les prétraitements (suppression des mots non porteurs de sens, lemmatisation, *etc.*) suffiront pour la tâche de segmentation thématique. Néanmoins, les systèmes exploitant la pondération des termes restent majoritaires et considèrent que la pondération donne un avantage dans le calcul de la cohésion lexicale. Par exemple, (Yaari, 1998) indique que la pondération de mots uniquement avec les

1. Pour rappel, un chunk est composé de plusieurs groupes de souffle et correspond à l'unité utilisée dans le calcul de la pondération des termes (identique à la notion de document en recherche d'information).

scores IDF de la pondération *TF-IDF* donne plus de crédibilité aux valeurs de la cohésion.

En ce qui nous concerne, nous avons fait des expériences évaluant l'impact de la pondération sur la tâche de segmentation. Pour cela, nous comparons l'algorithme de segmentation dans lequel aucune pondération spécifique n'est utilisée (*i.e.* uniquement à partir des fréquences des mots) avec deux autres versions pondérées : le système de segmentation avec pondération uniforme (voir la section 3.3.2) et un autre avec pondération *Oracle* (*i.e.* les chunks sont déterminés à partir des segments de référence). Les conditions *Oracle* nous permettent d'évaluer la performance maximale de notre système basé sur la pondération *intra-document*. En effet, avec la pondération *Oracle* chaque chunk représente un segment thématique et permet ainsi une pondération plus fiable que celle basée sur les chunks uniformes.
Les résultats du tableau 4.1 montrent que le système à pondération uniforme donne de meilleurs résultats qu'un système basé uniquement sur la fréquence des mots (*i.e.* sans pondération). Les résultats obtenus en conditions *Oracle* montrent

	Corpus MC-0813			*Corpus MCS7-14*		
	F-mesure	*CouvN*	*CouvD*	*F-mesure*	*CouvN*	*CouvD*
Sans pondération	58.1	34.1	39.8	57.0	38.6	52.0
Uniforme	59.4	44.2	53.8	58.4	44.3	52.8
Oracle	74.3	56.5	65.7	71.2	58.7	70.0

***Tableau 4.1** – Impact de la pondération sur la performance de notre système de segmentation.*

qu'il existe une importante marge de progression. Ce constat est encourageant et permet d'identifier les pistes d'amélioration. On peut notament constater que plus le début des chunks est proche des frontières thématiques, meilleure est la qualité de la segmentation. En ce sens, nous décrivons dans la section suivante deux approches pour améliorer la version de la pondération *uniforme* de (Malioutov et Barzilay, 2006).

4.3 Deux propositions de pondération intra-document à base de chunks

Nous proposons deux approches pour pondérer les termes de l'émission (Bouchekif *et al.*, 2014a) : la pondération basée sur des *informations structurelles* et la pondération *itérative*. Le principe général est de découper l'émission en

N chunks de *différentes tailles* où chaque chunk s'apparente à un document en recherche d'information. Idéalement, le début de chaque chunk correspond au lancement d'un nouveau thème.

4.3.1 Pondération basée sur des informations structurelles

Nous proposons d'utiliser des informations structurelles pour définir les chunks. Nous avons en particulier sélectionné l'information liée au présentateur principal de l'émission. Le principe général du système de détection du présentateur principal est le suivant :

- La première étape consiste à détecter les changements de locuteurs en déterminant les tours de parole pour chaque locuteur qui figure dans l'émission. Les tours de parole sont identifiés par une étiquette unique anonyme.
- Dans la deuxième étape, l'émission est découpée en N morceaux de même taille, dans notre cas le paramètre N est fixé à 10.
- Finalement, pour chaque locuteur, nous comptons le nombre de morceaux dans lequel il est présent. Ainsi, nous considérons le locuteur le plus présent dans les morceaux comme étant le présentateur principal de l'émission. Nous formulons le postulat que le présentateur principal est le locuteur présent de façon régulière tout au long de l'émission.

La fusion du présentateur principal et des indices lexicaux donnent de bons résultats pour la segmentation thématique (Bouchekif *et al.*, 2013a), (Charlet *et al.*, 2015a).
Dans ce travail, nous utilisons l'information liée à la détection du présentateur principal comme une première partition (segmentation initiale) qui servira ensuite dans le calcul de la pondération des mots de l'émission. Nous considérons que le début de chaque intervention du présentateur principal est le début d'un nouveau chunk (voir la figure 4.3).

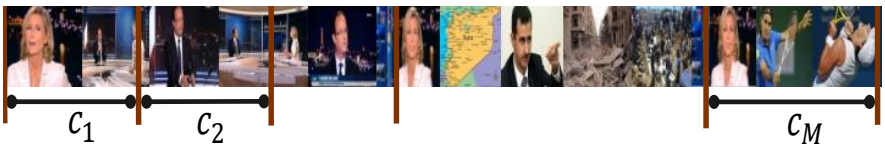


FIGURE 4.3 – *Découpage de l'émission en chunks avec l'approche structurelle.*

Il est intéressant de mentionner que la structure de l'émission et la performance de l'outil de détection du présentateur principal ont une influence directe sur notre approche. Voici quelques exemples où cette technique de pondération semble devenir moins efficace :

— Dans les zones où un journaliste enchaîne une série de thèmes sans illustration par des reportages (brèves).

— Dans les zones où le présentateur principal fait une interview avec un invité sur un thème donné.
— Lorsque l'outil peine à bien détecter le présentateur principal ou n'a pas déterminé correctement le nombre de présentateurs principaux (aucun, un seul ou plusieurs) ainsi que les zones correspondant à leurs interventions.

Remarque : la pondération basée sur des informations structurelles ne se restreint pas au présentateur principal. Elle peut être élargie à d'autres indicateurs de changement de thèmes comme l'apparition de titres incrustés.

4.3.2 Pondération itérative

Dans cette approche, les poids sont calculés de façon itérative. Chaque itération propose un découpage de l'émission en chunks à partir duquel les poids sont ré-estimés lors de la prochaine itération.

1. Initialement, l'émission est coupée en N chunks uniformes. Le nombre N est obtenu automatiquement pour chaque émission en divisant la durée de l'émission par la durée moyenne des segments thématiques (106 secondes) estimée sur le corpus de développement *MC-0813*. Les indices de chaque premier groupe de souffle de chaque chunk forment l'ensemble initial des frontières et sont placés dans le vecteur hyp_0.
2. À l'itération i :
   — **Pondération**
   Les hypothèses d'itération $i-1$ notées hyp_{i-1} sont utilisées pour estimer la pondération *TF-IDF (i)*.
   — **Segmentation thématique**
   L'algorithme de segmentation thématique est appliqué en prenant en compte la nouvelle matrice des poids dans le calcul de la cohésion lexicale entre les blocs adjacents. Le poids associé au terme t dans la représentation vectorielle d'un bloc b est alors donné par :

$$v(b,t) = \sum_{x \in b} \left(f_{(x,t)} \times w\left(c\left(x\right), t\right) \right) \tag{4.4}$$

   où : $f_{(x,t)}$ est la fréquence du terme t dans le groupe de souffle x.
   $w(c(x),t)$ est le poids cumulé du terme t dans le chunk x. Pour rappel, la formule de $w(c(x),t)$ est donnée dans la sous-section 3.3.2.

   La cohésion au niveau d'une frontière potentielle j entre les deux blocs b_j et b_{j+1} est donnée par $cohesion(j) = cosine(V_j, V_{j+1})$.
   où : $V_j = \sum_{t \in b_j} v(b_j, t)$

3. Les segments d'hypothèses obtenus sont présentés à l'étape 2.
   L'algorithme s'arrête lorsque la segmentation s'est stabilisée (*i.e* pas de changement significatif entre les hypothèses de deux itérations successives hyp_{i-1} et hyp_i).

Afin de mesurer objectivement cette stabilisation, nous utilisons la métrique p_k (voir la section 2.4.1) qui compare les deux segmentations : courante (hyp_i) et précédente (hyp_{i-1}). Nous avons fixé la valeur k à 6 groupes de souffle. L'algorithme s'arrête lorsque la valeur de p_k entre hyp_{i-1} et hyp_i est proche de 1 (*i.e.* $1 - p_k(hyp_{i-1}, hyp_i) < \epsilon$). Nous n'avons pas étudié la preuve de la convergence de l'algorithme, mais dans le cas où le critère d'arrêt n'est pas vérifié au bout de la $6^{ème}$ itération, l'algorithme s'arrête.

L'avantage de l'approche basée sur la pondération itérative est qu'elle ne nécessite aucune information sur la structure de l'émission. Néanmoins, dans le cas où l'émission est présentée par un présentateur principal, l'approche basée sur des informations structurelles donne généralement de meilleures performances.

4.4 Expériences et Résultats

La figure 4.4 illustre les performances du système au niveau de la détection des frontières pour le corpus de développement *MC-0813*, avec les quatre pondérations : *Uniforme*, *Itérative*, *Structurelle* et *Oracle*.

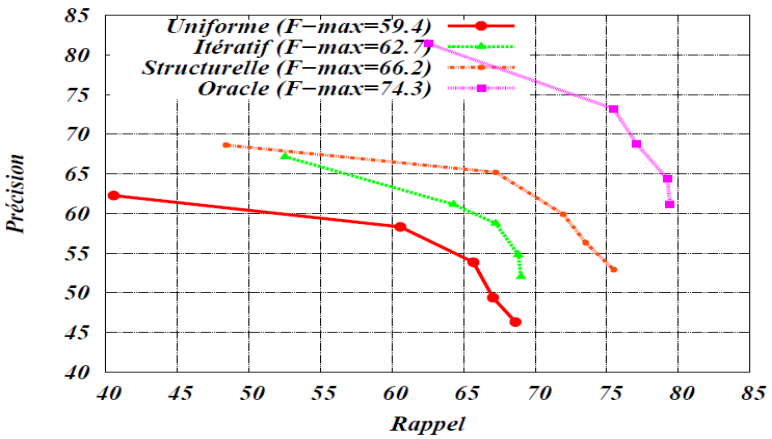


FIGURE 4.4 – *Impact de la stratégie de pondération sur la performance du système de segmentation avec le corpus MC-0813.*

Les courbes rappel/précision ont été obtenues en faisant varier le seuil de segmentation s_α qui prend des valeurs allant de 0.82 à 0.90. On constate que

la *F-mesure* est meilleure pour les systèmes utilisant la pondération itérative ou structurelle, par rapport à celui qui pondère les mots à l'aide des chunks uniformes. Le système de segmentation basé sur la pondération *Oracle* obtient les meilleurs résultats. Ces résultats confirment l'importance de la pondération dans la segmentation thématique.
Le tableau 4.2 détaille les performances du système à l'aide des trois métriques d'évaluations : *F-mesure*, *CouvD* et *CouvN* pour le point qui maximise la *F-mesure* de la figure 4.4. Notons que, le système de segmentation thématique obtient les meilleures performances en fixant s_α à 0.88.

Corpus *MC-0813*									
Condition de pondération	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Uniforme	60.6	58.3	59.4	45.6	42.8	44.2	53.8	53.8	53.8
Itératif	64.3	61.1	62.7	47.9	44.3	46.0	55.3	55.3	55.3
Structurelle	67.2	65.2	66.2	51.9	48.8	50.3	57.9	57.8	57.8
Oracle	75.5	73.2	74.3	58.2	54.9	56.5	66.0	65.4	65.7

Tableau 4.2 – *Performance du système avec les différentes stratégies de pondération sur le corpus MC-0813.*

Le système basé sur la pondération *itérative* améliore l'approche *uniforme*, en détectant 13 nouveaux segments corrects, permettant de passer d'une *CouvN* de 44.2% à 46.0%. Au niveau des frontières, un gain de 3.3 points de *F-mesure* a été enregistré en appliquant la pondération itérative.
L'utilisation des informations structurelles dans notre cas améliore la performance du système en détectant 36 nouveaux segments par rapport à l'approche *uniforme*. Ainsi la mesure *CouvD* augmente de 4% et *CouvN* de 6% en absolu. Les résultats du système avec une pondération *Oracle* nous montrent à nouveau l'importance de la pondération. En effet, une marge de progression importante est constatée au niveau des frontières et segments (14.9% et 12.3% respectivement) entre un système de segmentation thématique basé sur la pondération *Oracle* et un autre basé sur la pondération *uniforme*.
La différence de gain entre l'approche *Itérative* et *Structurelle* vient principalement de la structure des journaux d'information télévisés : le présentateur principal donne un aperçu du thème qui est ensuite détaillé par un reportage ou une interview. Deux éléments empêchent d'avoir une performance maximale : les interviews dans le plateau et les brèves.

Afin d'évaluer la robustesse de nos systèmes, nous relançons les mêmes expériences que précédemment sur le corpus de test *MCS7-14*. Certaines émissions de ce corpus présentent en effet quelques particularités. Par exemple, *EuroNews* ne dispose pas du tout de présentateur principal (succession de reportages) et *France_Journal7* est présenté par plusieurs présentateurs (au moins

deux). Par conséquent, notre système de segmentation basé sur la pondération structurelle devient moins efficace. Face à ce problème, nous avons mis en place une configuration nommée *Standard* qui associe à chaque émission télévisée le type de pondération adéquat. La configuration *Standard* est donnée dans le tableau 4.3. Celle-ci privilégie l'approche *Structurelle*. Lorsque la structure de l'émission ne permet pas son utilisation, c'est l'approche itérative qui est utilisée.

	Anchor	Itérative
ArteJournal	×	
D8_LeJT	×	
EuroNews		×
France2_Journal13Heures	×	
France2_Journal20Heures	×	
France2_Journal7Heures		×
France2_Journal8Heures	×	
France3_1213JournalNational	×	
France3_1920JournalNational	×	
M6_Le1245	×	
M6_Le1945	×	
NT1_Infos		×
TF1_13Heures	×	
TF1_20Heures	×	

***Tableau 4.3** – Configuration Standard : type de pondération utilisée par émission*

Les courbes *rappel/précision* de la figure 4.5 donnent les performances du système en terme de nombre de frontières correctement retournées. Pour chaque courbe, nous faisons varier le seuil de segmentation s_α. Pour évaluer plus en détails le système de segmentation au niveau des segments correctement retournées (*CouvN* et *CouvD*), nous fixons s_α à 0.88. Les résultats sont donnés dans le tableau 4.4.

Condition de pondération	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Uniforme	57.9	59.0	58.4	44.3	44.3	44.3	52.6	53.1	52.8
Itératif	64.3	61.2	62.7	48.3	45.9	47.1	56.2	56.5	56.4
Standard	70.7	68.1	69.4	58.5	55.2	56.8	68.4	68.4	68.4
Oracle	72.3	69.5	71.2	60.8	56.7	58.7	70.1	69.9	70.0

***Tableau 4.4** – Performance du système avec les différentes approches de pondération sur le corpus MCS7-14.*

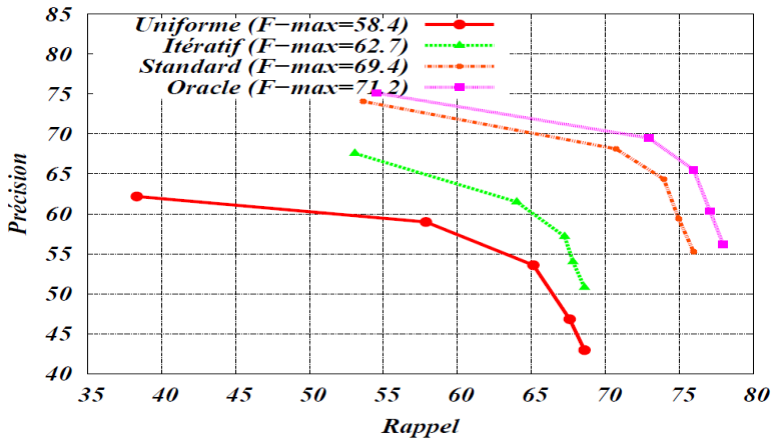


FIGURE 4.5 – *Impact de la stratégie de pondération sur la performance du système de segmentation avec le corpus MCS7-14.*

Les tendances observées sur le corpus *MC-0813* se reproduisent avec les émissions de *MCS7-14*, c'est-à-dire que la pondération définie avec les conditions *Standard* (très proche de la configuration structurelle) donne de meilleures performances par rapport à la pondération itérative. On observe également une importante marge de performances entre les résultats obtenus avec *Oracle* et ceux avec *Uniforme*. La seule différence par rapport aux résultats obtenus avec le corpus *MC-0813* est que la marge entre la configuration *Standard* et l'*Oracle* est moins importante sur le corpus *MCS7-14* (8.1 et 1.8 points F-mesure pour les corpus *MC-0813* et *MCS7-14* respectivement).

Pour comprendre cette réduction de marge entre les deux corpus, nous avons mené des expériences complémentaires. Cette fois-ci, nous considérons le début de chaque intervention du présentateur principal directement comme une frontière thématique. Pour le corpus *MC-0813*, l'évaluation au niveau des frontières donne un *R*=63.7, une *P*=60.0 ce qui donne une *F-mesure*=61.8. Avec le corpus *MCS7-14*, le système obtient une F-mesure qui vaut 66.7% (*R*=74.5, *P*=60.4). Nous déduisons de ces résultats que les interventions du présentateur principal coïncident souvent avec un changement de thèmes sur les deux corpus. Cependant, ce constat est beaucoup plus marquant dans le corpus le *MCS7-14* que dans *MC-0813*. En d'autres termes, les chunks fournis par l'approche *Structurelle* sur *MCS7-14* sont plus proches des frontières thématiques de référence. La pondération est donc meilleure. Notons que dans cette évaluation, les émissions d'Euronews, France2_Journal7Heures et NT1 sont exclues en raison de l'absence ou de la présence de plusieurs présentateurs principaux.

Nous nous intéressons maintenant à l'évaluation du système par rapport à la nature des segments. Pour rappel, nous avons considéré deux types de segments : *court* et *long* (voir la section 2.3.2). Le tableau 4.5 donne les performances

du système en terme de nombre de segments correctement retournés. Indé-

	Corpus MC-0813				*Corpus MCS7-14*			
	Seg. Longs		Seg. Courts		Seg. Longs		Seg. Courts	
	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$
Uniforme	52.7	58.3	14.3	8.0	51.8	55.2	20.3	16.9
Itératif	55.5	59.3	14.3	8.3	56.5	58.7	22.0	16.4
Standard	59.1	62.5	20.0	12.6	68.3	69.5	26.7	20.5
Oracle	67.3	70.6	18.1	11.7	71.2	71.4	27.1	20.6

***Tableau 4.5** – Influence de la taille des segments sur la qualité de segmentation avec les différents approches de pondération.*

pendamment de l'approche, le système est performant sur des thèmes longs. Cependant, un très faible nombre de segments courts ont été trouvés. Malgré cela, nos approches améliorent non seulement les segments longs mais aussi les courts. Toutefois, ces derniers restent toujours difficilement détectables.

Enfin, nous validons l'apport de la pondération *intra-document* avec le corpus *MCS5-15*, les résultats sont donnés dans le tableau 4.6. Sans surprise, la hiérarchie des performances est respectée. C'est-à-dire que l'approche *Itérative* reste plus performante que l'approche *Uniforme* et celle *structurelle* reste plus performante que *l'itérative*.

Condition de	Nbre frontières			Nbre segments			Durée segments		
pondération	*R*	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Uniforme	57.5	56.0	56.7	45.8	44.9	45.3	56.3	57.4	56.8
Itératif	60.2	57.1	58.6	46.5	43.9	45.1	56.4	57.1	56.8
Structurelle	65.3	60.8	63.0	49.9	45.3	47.4	61.0	60.9	60.9
Oracle	74.9	72.9	73.9	59.3	58.2	59.0	70.0	69.8	69.9

***Tableau 4.6** – Performance du système avec les différentes approches de pondération sur le corpus MCS5-15.*

Nous constatons aussi que notre système de segmentation est moins performant sur le corpus *MCS5-15*. Cela peut s'expliquer par les particularités des thèmes constituant les émissions de ce corpus. En effet, la période du 26 au 27 janvier 2015 coïncide avec des événements marquants comme les élections législatives en Grèce, le crash d'un avion militaire en Espagne, la commémoration d'Auschwitz, *etc.* Chaque événement n'est pas traité dans un seul segment thématique mais est découpé au sein de plusieurs thèmes, chacun abordant une information particulière. Cette multiplicité d'informations portant sur un même événement rend la tâche de la segmentation thématique beaucoup plus difficile, puisque les différents segments restent liés les uns aux autres. Par exemple,

l'événement "les élections Grecques" est présenté en six thèmes (voir le tableau 1.1). Or, ces derniers ne sont pas totalement déconnectés les uns aux autres.

4.5 Conclusion

Nous venons d'étudier l'influence de la pondération dans un système de segmentation basé sur le calcul de la cohésion lexicale. Les deux approches proposées (avec des informations structurelles et itératives) améliorent nettement la performance du système de segmentation thématique. L'utilisation du présentateur principal dans la pondération donne généralement de meilleurs résultats que l'approche itérative. Néanmoins, cette dernière a comme avantage de pouvoir être appliquée sur n'importe quelle émission (généricité) car elle est indépendante de la structure de l'émission. De l'analyse des résultats, nous avons observé que la majorité des segments non détectés concerne surtout les thèmes traitant de thématiques proches (deux reportages consécutifs portant sur un même pays) et les thèmes ayant une faible répétition de mots.

Chapitre 5

De la cohésion lexicale à la cohésion de la parole

Sommaire

5.1 Introduction

Le signal audio d'une émission contient des informations autres que les mots prononcés, comme les locuteurs présents dans le journal, les événements sonores, *etc.* La prise en compte de ces informations peut améliorer la qualité de la segmentation thématique en particulier dans les parties où il y a une faible répétition des mots. En effet, dans ces zones marquées par un manque de redondance des mots, les systèmes de segmentation thématique basés uniquement sur la cohésion lexicale se révèlent parfois impuissants à détecter des changements de thèmes. Il peut arriver également que les systèmes de segmentation

indiquent des ruptures thématiques alors qu'en réalité elles n'existent pas. Nous pouvons citer les deux cas suivants :

— 1^{er} cas : une fausse alarme au milieu d'un reportage (voir la $i^{ème}$ frontière potentielle de la figure 5.1).

— $2^{ème}$ cas : une fausse alarme avant même l'achèvement d'une interview (voir la $j^{ème}$ frontière potentielle de la figure 5.1).

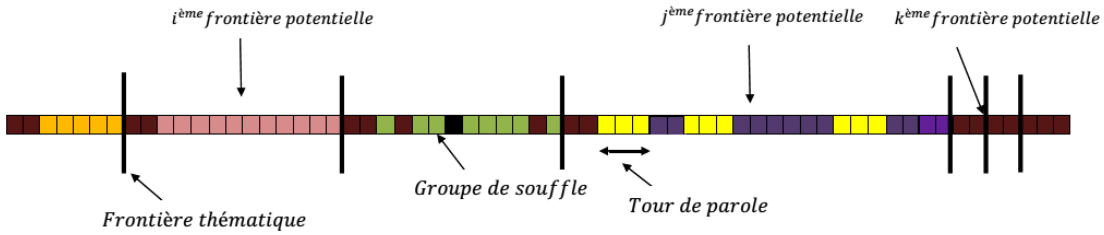


FIGURE 5.1 – *Corrélation entre la distribution des locuteurs et les changements de thèmes.*

L'exploitation conjointe de la distribution des mots et des locuteurs peut empêcher ce genre de fausses alarmes et même permettre de détecter de nouvelles frontières. Les systèmes de structuration en locuteurs rendent les informations liées aux locuteurs accessibles et exploitables. De ce fait, notre système est capable de traiter n'importe quelle émission sans aucune information *a priori*.

Nous proposons d'intégrer la répartition des mots et des locuteurs au moment du calcul de la similarité entre les différentes parties de l'émission. La mesure englobant les deux distributions est alors appelée *cohésion de parole*. En s'appuyant sur ce principe, une frontière potentielle est validée lorsque la distribution conjointe des mots et des locuteurs diffère suffisamment de part et d'autre de la frontière.

Les algorithmes de segmentation exploitant l'information liée aux locuteurs le font essentiellement dans un cadre supervisé (voir la section 1.4.2). Dans ce chapitre, nous proposons une approche non supervisée qui intègre la distribution des mots et des locuteurs. L'avantage de notre approche est qu'elle n'a aucune influence sur la généricité du système.

5.2 Introduction d'un nouveau paradigme : la cohésion de la parole

Pour rappel, le calcul de la cohésion lexicale est basé sur les mots présents plusieurs fois dans un même thème. Les mots non discriminants sont soit supprimés à l'aide de l'étape de pré-traitements, soit pénalisés par le processus de

pondération. Les mêmes tendances de la distribution des mots peuvent se produire avec la distribution des locuteurs. En effet, si le présentateur principal de l'émission intervient généralement tout au long du journal télévisé (JT), il est plus probable que les autres locuteurs comme les reporters ou les invités n'interviennent que dans un seul thème. La notion de cohésion, appliquée sur la distribution des mots, peut s'étendre à la distribution des locuteurs. Pour obtenir la distribution des locuteurs, le signal audio de chaque émission est structuré en locuteurs. Ceci permet d'obtenir ou des segments n'appartenant qu'à un seul locuteur, ou des segments de silence. La cohésion lexicale et la cohésion en locuteurs sont ainsi regroupées en une seule notion : la cohésion de la parole (*speech cohesion*).

5.2.1 Structuration en locuteurs

La structuration en locuteurs (en anglais *speaker diarization*) consiste à découper un flux sonore en segments contenant les énoncés oraux d'un seul locuteur. Aucune information *a priori* sur les locuteurs du document n'est disponible (comme le nombre de locuteurs, les modèles acoustiques des locuteurs présents dans l'émission, *etc.*) Le système utilisé dans ce travail effectue la tâche de structuration en locuteurs en deux étapes : la *segmentation* et le *regroupement hiérarchique ascendant*. Toutes les deux sont basées sur le critère *BIC* (*Bayesian Information Criterion*). L'étape de segmentation consiste à identifier les points de changement de locuteurs ainsi que le type de chaque segment (parole, silence, bruit, musique). Le regroupement permet de mettre une étiquette unique (de type *locuteur A*, *locuteur B*) pour les segments ayant les mêmes caractéristiques. La première étape de la structuration en locuteurs permet d'obtenir une segmentation initiale dans laquelle les clusters sont assez purs et contiennent suffisamment de données pour permettre de modéliser le locuteur par un mélange de gaussiennes. Ensuite, avec une telle modélisation, un processus itératif de segmentation *via* un décodage de *Viterbi*, et de regroupement par le critère *GLR* (Generalized Likelihood Ratio) (Barras *et al.*, 2006) est réalisé. La mise en œuvre et les performances du système sur des données de type *Broadcast News* sont données dans (Charlet *et al.*, 2013).

Les trois corpus *MC-0813*, *MCS7-14* et *MCS5-15* contiennent respectivement 1496, 2773 et 966 locuteurs différents. La figure 5.2 donne la répartition des locteurs par rapport au nombre de leurs interventions dans différents thèmes, durant l'émission. Les chiffres ont été obtenus avec notre système de structuration en locuteurs.
En analysant les histogrammes, nous constatons que les trois corpus contiennent plus de 70% de locuteurs intervenant uniquement dans un seul thème d'une émission et 16% de locuteurs sont présents dans deux thèmes différents. Le

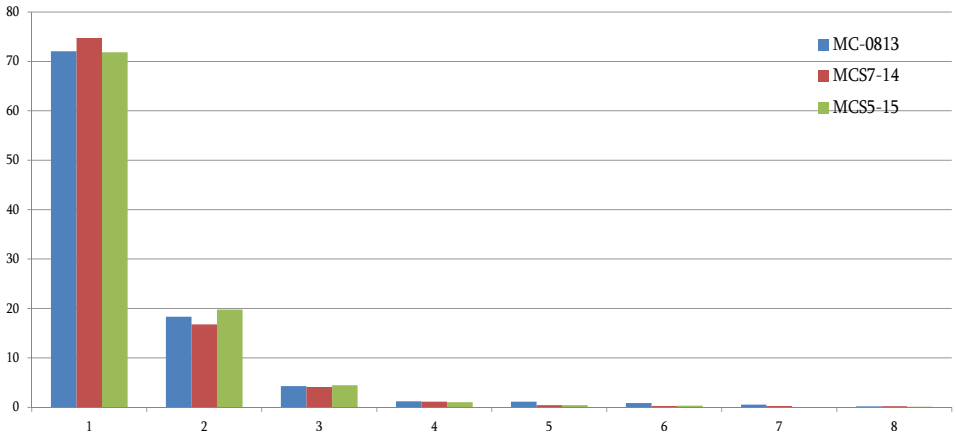


FIGURE 5.2 – *Répartition des locuteurs par leurs nombres d'interventions dans des thèmes différents durant l'émission.*

présentateur principal est le seul qui a plus de trois interventions dans une même émission. Il faut noter que la segmentation et le regroupement réalisés manuellement sur 33 émissions du corpus *MC-0813* donnent 96.4% de locuteurs intervenant dans un seul thème. Si nous faisons la même expérience, mais cette fois-ci, avec notre système de structuration en locuteurs, nous obtenons 70.2% de locuteurs qui n'interviennent dans un seul thème. Il est clair que les erreurs du système ont une mauvaise influence sur la segmentation thématique produite. Ainsi, plus le système de structuration en locuteurs aura de bonnes performances, plus il sera à même d'aider le système de segmentation thématique.

5.2.2 Cohésion de la parole

L'idée de la cohésion de la parole est de considérer l'identifiant du locuteur comme un terme, c'est-à-dire que chaque groupe de souffle est représenté non seulement par les mots prononcés mais également par l'indice du locuteur correspondant. Pour cela, nous procédons préalablement à l'étape de synchronisation entre les groupes de souffle déterminés par le système de transcription et les tours de parole obtenus par le système de structuration en locuteurs.

La figure 5.3 est un exemple issus d'un fichier que nous présentons en entrée de notre système de segmentation.

Chaque ligne représente un groupe de souffle, son instant de début et de fin sont représentés par les colonnes 1 et 2, la troisième colonne correspond à l'indice du locuteur, le reste contient les mots qui ont été prononcés (après les pré-traitements). Un groupe de souffle vide signifie que les mots prononcés sont

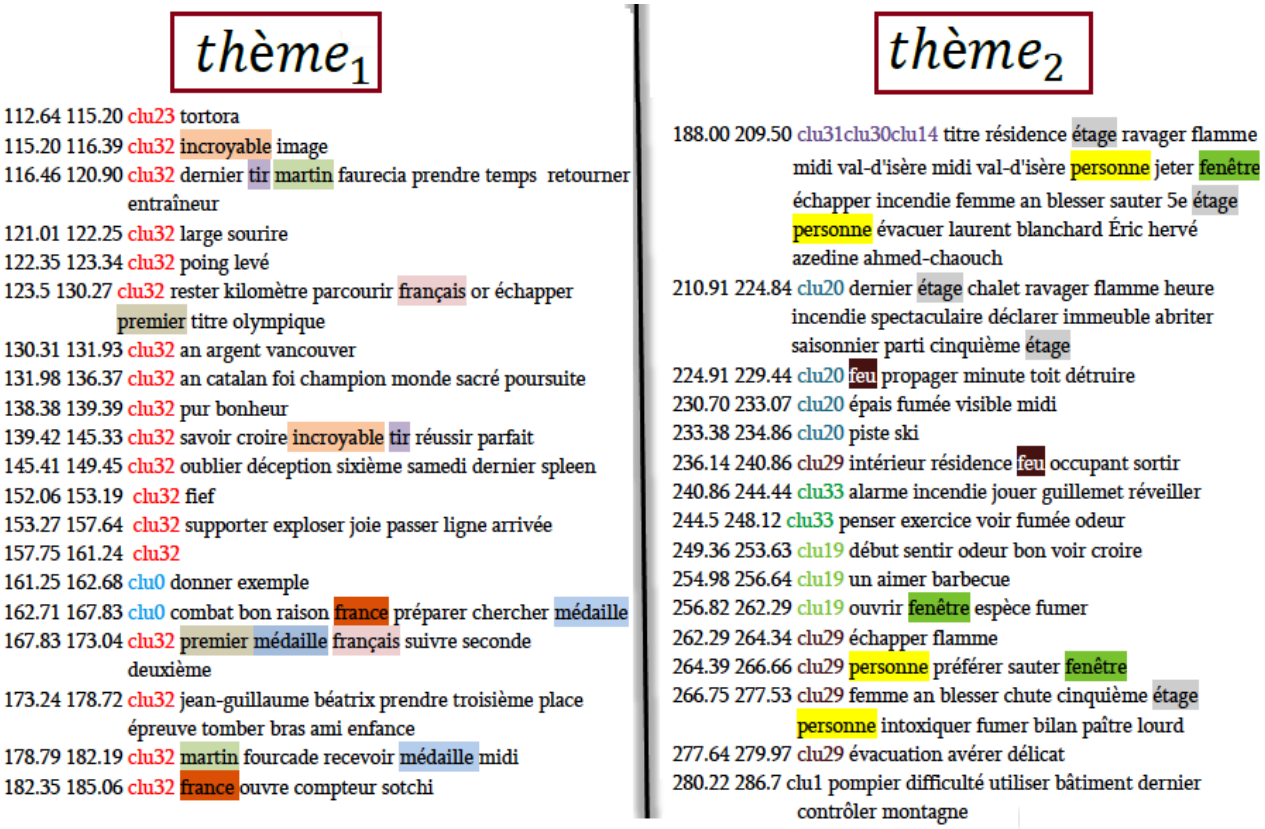


FIGURE 5.3 – *Représentation conjointe de la distribution lexicale et des locuteurs.*

filtrés par leur score de confiance et/ou par leur catégorie grammaticale (pour plus de détails voir la section 1.3.1).

Nous pouvons voir la complémentarité entre les deux distributions. En effet, dans le premier thème (la partie gauche de la figure 5.3), la redondance des mots n'est pas suffisante. La répétition de certains mots importants ne couvre pas la totalité du document. C'est le cas pour les mots *médaille* et *france* se trouvant vers la fin du thème. Par ailleurs, on observe une forte redondance du locuteur portant l'identifiant *clu32*. La prise en compte conjointe de ces deux distributions peut empêcher la production de fausses alarmes.

Prenons maintenant le deuxième thème (la partie droite de la figure 5.3), certains mots importants comme *personne*, *étage* et *fenêtre* apparaissent du début à la fin du document. Cependant, la redondance au niveau des locuteurs reste insuffisante pour déterminer la cohésion thématique (plusieurs intervenants et chacun d'eux ne prend la parole qu'une seule fois). La combinaison de ces deux informations est un atout important pour le système de segmentation.

5.3 Intégration de la distribution des locuteurs dans le calcul de la cohésion lexicale

Dans notre algorithme de segmentation, l'émission est représentée sous la forme d'une matrice de taille $n \times m$ où n est la taille du vocabulaire et m est le nombre de groupes de souffle. L'élément $e_{i,j}$ représente le nombre d'occurrences du $i^{\text{ème}}$ terme dans le $j^{\text{ème}}$ groupe de souffle. Le vecteur de représentation associé à un groupe de souffle était jusqu'ici constitué d'autant de composantes que de termes prononcés dans l'émission. Ce vecteur est désormais augmenté d'autant d'éléments que le JT contient de locuteurs différents (voir la figure 5.4). Cette représentation permet de faire cohabiter la distribution des mots et des locuteurs dans le calcul de la cohésion.

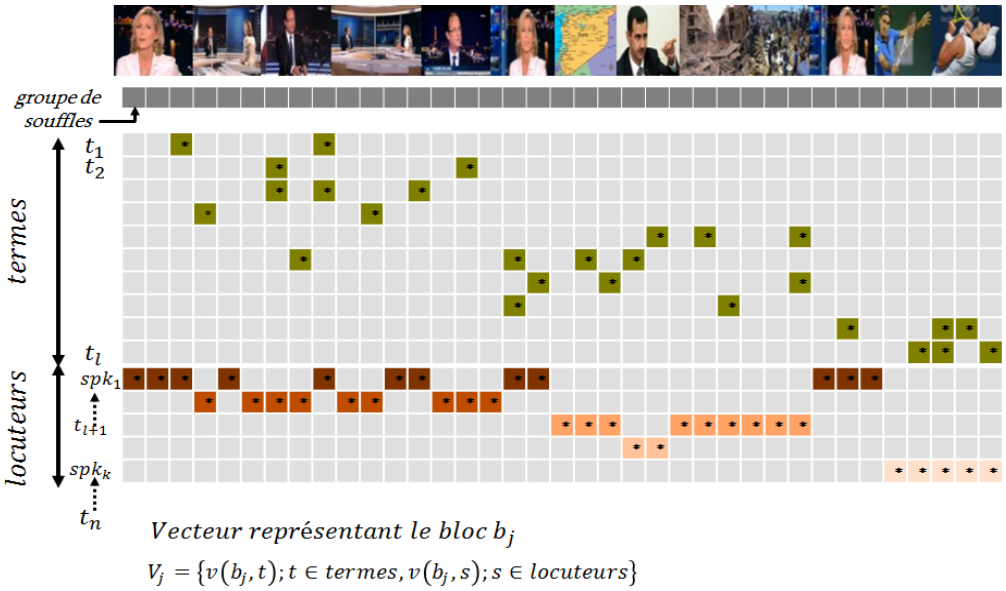


FIGURE 5.4 – *Intégration de la distribution dans la représentation vectorielle du document.*

Un simple comptage du nombre de mots pour le locuteur associé au groupe de souffle n'est pas approprié car il génère une contribution trop importante de l'information liée au locuteur à l'échelle de la fenêtre sur laquelle est calculée la cohésion. Afin d'équilibrer la contribution des mots et des locuteurs, nous avons essayé d'effectuer une normalisation relativement au nombre maximum de mots (max) dans l'ensemble des groupes de souffle du journal télévisé (Bouchekif *et al.*, 2014b). Cette méthode améliore les résultats par rapport à une valeur de 1.0 (*i.e* avec un simple comptage). Cependant, une valeur empirique de 0.5 donne de meilleurs résultats, et c'est cette valeur que nous utiliserons dans nos expériences.

Le principe de la pondération appliqué sur les mots reste le même pour les

étiquettes de locuteurs (voir la section 4.3). Pour ces derniers, la pondération permettra de pénaliser les locuteurs présents durant toute l'émission (comme le présentateur principal par exemple) et de favoriser les locuteurs qui interviennent localement pour un thème donné. Le poids associé au locuteur l dans la représentation vectorielle d'un bloc b est une valeur pondérée $v(b,l)$ donnée par :

$$v(b,l) = \sum_{x \in b} \left(f_{(x,l)} \times w\left(c\left(x\right), l\right) \right) \tag{5.1}$$

où $w(c(x),l)$ est le poids cumulé du terme l dans le chunk x (le terme l correspond à l'étiquette de locuteur l).
$f_{(x,l)}$ est la valeur associée au locuteur l dans le groupe de souffle x.

L'avantage de cette approche est que la mesure de cohésion prend en compte la redondance au niveau des mots et des locuteurs.

5.4 Expériences et résultats

5.4.1 Impact de la cohésion de la parole

La figure 5.5 illustre les performances du système sans et avec prise en compte de la distribution de locuteurs. Les courbes *Rappel/Précision* ont été obtenues en faisant varier le seuil de segmentation thématique S_α. Le système de segmen-

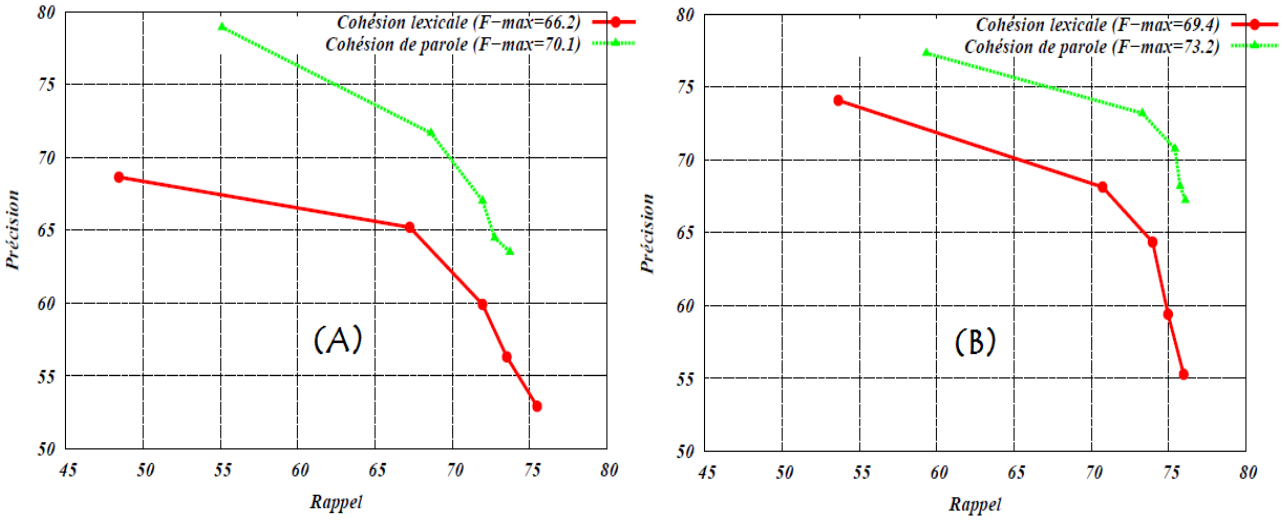


FIGURE 5.5 – *Performances du système de segmentation avec la cohésion lexicale et de la parole pour les corpus MC-0813 (A) et MC-S714 (B).*

tation obtient de meilleurs résultats en tenant compte des indices de locuteurs

dans le calcul de la cohésion. En effet, la prise en compte conjointe de la distribution des mots et des locuteurs augmente la qualité de la segmentation thématique. On peut le voir grâce au gain observé pour les deux corpus au niveau des frontières et des segments. Pour le corpus *MC-0813*, la *F-mesure* passe de 66.2 à 70.1 et de 69.4 à 73.2 pour le corpus *MCS7-14*.
Le tableau 5.1 donne les performances du système avec les trois métriques d'évaluations (*F-mesure*, *CouvD* et *CouvN*) pour le point qui maximise la *F-mesure* de la figure 4.4. L'intégration de la distribution de locuteurs augmente les mé-

Type de cohésion	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CouvD*
Corpus *MCS7-14*									
Lexicale	70.7	68.1	69.4	58.5	55.2	56.8	68.4	68.4	68.4
Parole	73.3	73.2	73.2	64.0	63.0	63.5	74.0	74.0	74.0
Corpus *MC-0813*									
Lexicale	67.2	65.2	66.2	51.9	48.8	50.3	57.9	57.8	57.8
Parole	68.6	72.0	70.1	58.9	59.9	59.4	69.4	68.5	69.0

***Tableau 5.1** – Performances du système basé sur la cohésion lexicale et la cohésion de la parole sur les deux corpus MC-0813 et MCS7-14.*

triques *CouvN* et *CouvD*. En effet, sur le corpus *MCS7-14*, le taux de segments correctement retournés *CouvN* passe de 56.8 à 63.5 et de 68.4 à 74.0 pour *CouvD* respectivement. Des tendances similaires ont été obtenues sur le corpus *MC-0813*.

Après analyse des résultats, nous constatons que l'insertion des étiquettes de locuteurs dans le calcul de la cohésion permet d'affiner la segmentation même si le JT contient peu de répétitions de mots. En effet, il peut arriver que notre système de segmentation propose une frontière avant même l'achèvement d'une intervention, voire au milieu d'un reportage au moment où un intervenant prend la parole. Dans ce cas, si le reporter reprend ultérieurement la parole, la redondance au niveau des locuteurs renforce la cohésion, ce qui rend les deux blocs adjacents liés thématiquement.

Nous constatons aussi que l'intégration de la distribution des locuteurs permet de limiter quelques phénomènes comme celui des *plateaux* : des faibles valeurs de cohésion durant longtemps sont observées. Ces phénomènes sont produits à cause du manque de répétition de mots ayant la même forme écrite. La figure 5.6 représente les courbes de la cohésion lexicale (a) et de la parole (b). La redondance au niveau des locuteurs a un effet positif non seulement pour supprimer les fausses alarmes (la $1^{ère}$ et la $3^{ème}$ hypothèse de la courbe *(a)*) mais aussi pour mieux placer les frontières ($3^{ème}$ hypothèse de la courbe *(b)*). Nous avons validé l'apport de la distribution des locuteurs sur le corpus *MCS5-15*.

Les résultats obtenus sont illustrés dans le tableau 5.2. Sur ces données, la prise en compte de la distribution du locuteur porte un gain absolu de 4.7%, 8.8% et 5.7% pour respectivement les mesures *F-mesure*, *CouvN* et *CouvD*.

Type de cohésion	*Corpus MCS5-15*		
	F-mesure	*CouvN*	*CouvD*
Lexicale	61.2	48.1	57.64
Parole	65.9	56.9	63.3

Tableau 5.2 *– Performances du système basé sur la cohésion lexicale et cohésion de la parole sur les données MCS5-15*

.

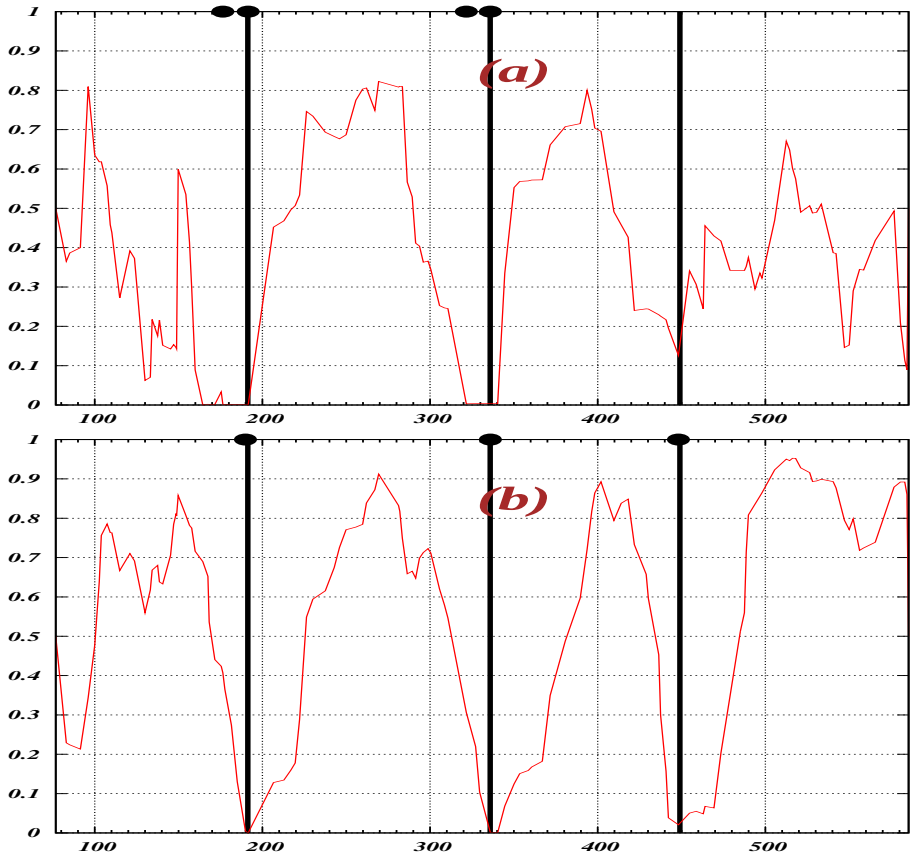


FIGURE 5.6 *– (a) courbe de la cohésion lexicale, (b) courbe de la cohésion de la parole. Les lignes verticales correspondent aux changements de thèmes, les petits ovales noirs représentent les frontières d'hypothèses.*

5.4.2 Influence de la taille des segments

Nous nous intéressons maintenant à l'évaluation du système au niveau des segments. Pour rappel, nous avons considéré deux types de segments courts et longs (voir la section 2.3.2). Les résultats sont exprimés par les deux mesures $\mathcal{R}_\mathcal{N}$ et $\mathcal{P}_\mathcal{N}$. Avec un système de segmentation thématique basé uniquement sur la cohésion lexicale, le système arrive à mieux détecter les segments longs. Un nombre limité de segments courts ont toutefois été bien détectés.

Type de cohésion	*Corpus MC-0813*				*Corpus MCS7-14*			
	Seg. Longs		Seg. Courts		Seg. Longs		Seg. Courts	
	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$
Lexicale	59.1	62.5	20.0	12.6	68.3	69.5	26.7	20.4
Parole	67.7	74.3	20.0	15.3	75.6	75.4	27.3	25.0

Tableau 5.3 – *Influence de la taille des segments sur la qualité de la segmentation.*

En intégrant la distribution de mots et de locuteurs, on constate, que pour les longs segments, le *Rappel* passe de 59.1% à 67.7% et de 68.3% à 75.6% pour respectivement les corpus *MC-0813* et *MCS7-14*. Une légère amélioration a été constatée au niveau des segments courts. En ce qui concerne la précision, la prise en compte de la répétition des locuteurs permet de supprimer quelques fausses alarmes, ceci se traduit par l'augmentation de la mesure $\mathcal{P}_\mathcal{N}$. Ces résultats montrent que la détection des segments courts par notre système de segmentation reste problématique.

De l'analyse des résultats obtenus, nous constatons que l'enchaînement de segments courts n'est pas favorable à notre système fondé sur la cohésion de la parole. En effet, des thèmes de moins de 30 secondes correspondent généralement à des brèves lues par le présentateur principal ou par un autre journaliste, impliquant des valeurs assez élevées de la cohésion de la parole. Ce constat est illustré dans la figure 5.7. Par exemple dans la période allant de 250 à 300, la cohésion de la parole prend des valeurs élevées alors qu'il y a peu de répétitions de mots du vocabulaire. On remarquera par ailleurs que les valeurs de la cohésion lexicale sont faibles (voir la courbe *(a)*). Ce phénomène met en évidence peu de changements de thèmes.

La prise en compte conjointe de la distribution de mots et de locuteurs permet essentiellement de limiter le nombre de fausses alarmes dans les zones souffrant de manque de répétitions de mots. En effet, le manque de répétitions donne de faibles valeurs de cohésion et conduit notre système à poser des frontières thématiques. Ce phénomène peut être corrigé avec la prise en compte de la redondance au niveau des locuteurs.

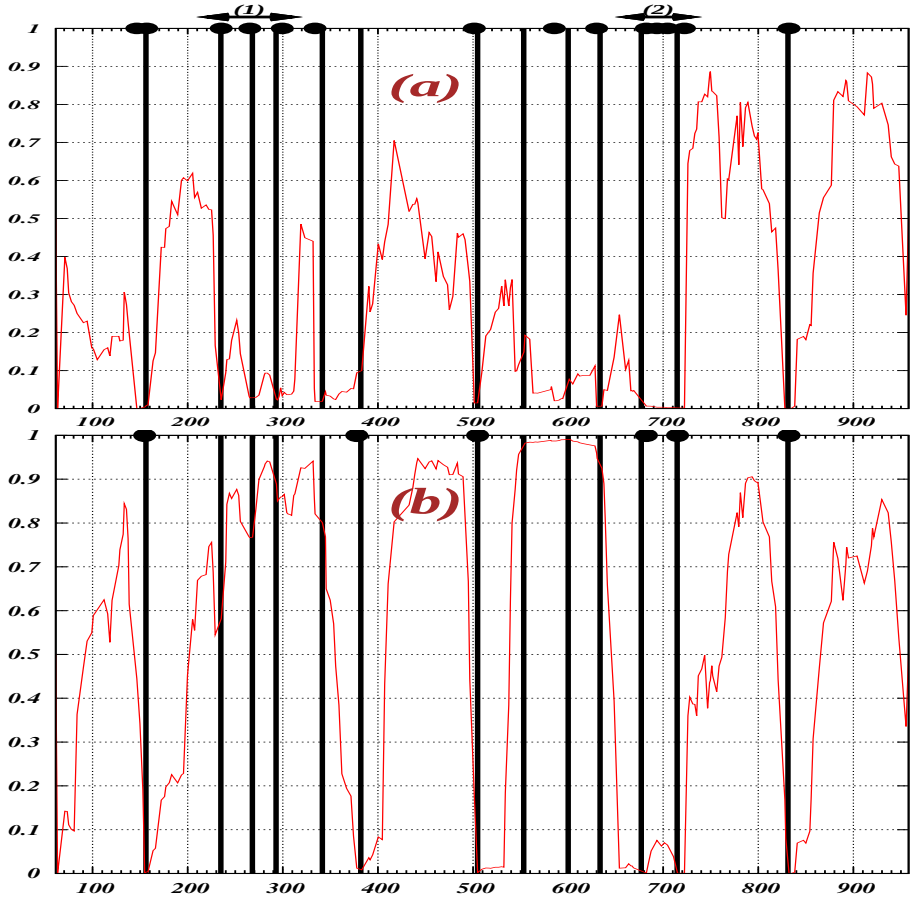


FIGURE 5.7 – *(a) courbe de la cohésion lexicale, (b) courbe de la cohésion de la parole. Les lignes verticales correspondent aux changements de thèmes, les petits ovales noirs représentent les frontières d'hypothèses.*

5.5 Exploitation de l'identification nommée des locuteurs

Les intervenants sont le plus souvent nommés, introduits par le présentateur et cités après leurs interventions ou à l'aide d'incrustation de textes (voir la figure 5.8). Dans ce cas, la cohésion de la parole peut être renforcée en remplaçant les labels des locuteurs (*clu1, clu2, etc.*) par leurs nom et prénom (*e.g.* François_HOLLANDE, Nicolas_SARKOZY). Ces derniers sont insérés artificiellement à la fin du groupe de souffle dans le cas où leur nom est prononcé (voir l'exemple illustré dans le tableau 5.4). L'utilisation de cette information pourra consolider le calcul de la cohésion en établissant un lien entre *ce qui est dit* et *qui le dit*.
Les segments thématiques longs sont des informations présentées soit sous forme

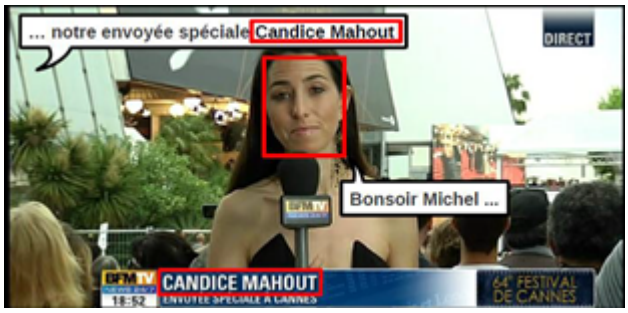


FIGURE 5.8 – *Sources d'information pour l'identification de personnes dans une émission télévisée (Béchet* et al.*, 2015).*

d'interviews soit de reportages. Dans les deux cas, il y a une forte possibilité d'avoir une redondance au niveau des locuteurs. Ceci rend d'autant plus visible l'importance de tenir compte de la distribution des locuteurs.

112.06 123.13 Gilles_BOULEAU ... ... notre envoyé milano dupont Milano_DUPONT
123.13 126.03 Milano_DUPONT ... ...
126.03 136.14 Milano_DUPONT ... ...
136.14 149.41 Milano_DUPONT ... ...
149.41 152.03 Milano_DUPONT ... ...
152.03 169.21 Gilles_BOULEAU merci dupont pour ces informations Milano_DUPONT

Tableau 5.4 – *Intégration des identités nommées des locuteurs dans la représentation vectorielle des documents.*

L'identification nommée peut être effectuée à partir de la transcription de la parole (Jousse, 2011) ou par un système multimodal (Favre *et al.*, 2013). Dans (Bouchekif *et al.*, 2014b), nous avons utilisé des émissions issues du projet *REPÈRE* [1]. Les locuteurs sont identifiés par le système présenté dans (Favre *et al.*, 2013). Pour voir l'impact de la prise en compte des noms de locuteurs dans la segmentation thématique, nous avons utilisé les sorties du système *REPÈRE* [2]. Les résultats obtenus dans (Bouchekif *et al.*, 2014b) montrent que l'identification nommée des locuteurs apporte une contribution favorable dans le calcul de la cohésion de la parole en augmentant la répétition des termes ayant la même forme écrite.

Remarque : dans (Bouchekif *et al.*, 2014b), nous avons travaillé sur les données *REPERE* (notamment les émissions *BFM Story* et *LCP INFO*) pour étudier

1. Le but du projet est d'identifier des individus apparaissant dans au moins une des modalités portées par l'enregistrement audiovisuel, qu'il s'agisse de locuteurs audibles ou de visages visibles à l'écran (Béchet *et al.*, 2015).
2. Merci à tous les collaborateurs du défi *REPÈRE*

le comportement de notre système de segmentation en prenant en compte les noms de locuteurs. Cependant, nous n'avons pas pu mettre en œuvre le système de reconnaissance de personnes sur nos corpus d'étude. Par conséquent, nous ne sommes pas en mesure de reproduire les mêmes expériences.

5.6 Conclusion

Dans ce chapitre, nous proposons de donner deux dimensions aux valeurs de similarité : la distribution des mots et la distribution des locuteurs. L'utilisation simultanée de ces deux informations améliore nettement les résultats en particulier pour détecter les segments longs. En revanche, les performances du système restent à améliorer lorsqu'il s'agit de segments courts. Ces derniers devraient faire l'objet d'un traitement spécifique. Le prochain chapitre poursuit l'enrichissement de l'espace vectoriel de documents mais cette fois-ci en intégrant les relations sémantiques entre les mots.

De nombreuses pistes restent à étudier pour enrichir la représentation vectorielle de l'émission. En effet, la façon d'intégrer les locuteurs que nous avons proposée peut s'étendre à d'autres sources d'informations permettant d'augmenter la redondance des termes ayant la même forme écrite. Ces informations doivent être discriminantes et prendre un nombre fini de valeurs (pour garantir la répétition). Par exemple, le domaine thématique des segments (sport, économie, international, *etc.*) est une information qui peut contribuer dans la reproduction de la répétition des termes. Les distances entre les groupes de souffle et les domaines thématiques peuvent être calculées en suivant la méthode proposée dans (Sakahara *et al.*, 2014).

Chapitre 6

Utilisation de relations sémantiques

Sommaire

6.1 Introduction

La plupart des systèmes de segmentation thématique actuels se fondent sur la redondance des termes ayant la même forme écrite (*i.e.* sans tenir compte des

liens sémantiques entre les mots). En s'appuyant sur ce principe, les valeurs de la cohésion lexicale ne reflètent pas parfaitement le degré de rapprochement thématique entre les différentes parties de l'émission. En prenant l'exemple ci-dessous, la similarité entre la phrase 1 et 2 est nulle alors que les deux phrases traitent du même thème « le conflit syrien ».

Phrase1 En syrie, les combats se poursuivent : au moins 70 corps de jihadistes ayant été rapatriés dans un hôpital.

Phrase2 Bachar El-Assad appelle au cessez le feu. L'armée syrienne contrôle environ 50% de la ville de Damas.

Pour résoudre ce problème, les linguistes mettent à notre disposition des ressources linguistiques comme les dictionnaires et les thesaurus (*WordNet* par exemple). Ce genre de ressources permet d'intégrer uniquement les relations de type synonymie (voiture, automobile) et hyperonymie (canich, chien). Idéalement, la cohésion doit prendre en compte les relations entre les mots proches thématiquement comme (Bachar, Syrie) ; (Bachar, El-Assad) ; (Syrie, Damas) ; (Syrie, Jihadiste) ; *etc.*

Les ressources traitant l'actualité peuvent satisfaire nos besoins. À partir d'un corpus d'articles multi-domaines, les mots apparaissant le plus souvent ensemble sont considérés comme proches thématiquement. Cela soulève plusieurs questions : comment estimer les relations sémantiques entre les mots ? Depuis quelles ressources ? Et comment intégrer la distance entre les mots dans le processus de segmentation thématique ?

Dans ce chapitre, nous tentons de répondre à ces questions, en commençant par définir les distances entre les mots les plus utilisées. Ensuite, nous donnons un aperçu rapide sur les méthodes qui prennent en compte les matrices de relations entre les mots dans le processus de segmentation thématique. Puis, nous présentons les données que nous avons utilisées dans l'extraction de relations sémantiques. Nous décrivons également la méthode que nous avons proposée pour intégrer les liens sémantiques entre mots dans le calcul de la similarité. Enfin, nous exposons les performances de notre système de segmentation thématique, cette fois-ci en tenant compte des relations sémantiques entre les mots.

6.2 Distance sémantique entre les mots

Dans la littérature, plusieurs méthodes d'extraction de relations sémantiques entre les mots ont été proposées. Elles peuvent être divisées en deux catégories :

1. Les mesures basées sur la théorie de l'information comme *PMI* (Pointwise Mutual Information) (Church et Hanks, 1989) et *NWD* (Normalized Web Distance) de (Vitányi *et al.*, 2009).
2. Les modèles basés sur le modèle vectoriel (Vector Space Model) comme

LSA (Latent semantic analysis) (Deerwester *et al.*, 1990), *Word2vec* (Mikolov *et al.*, 2013a) et *Glove* (Pennington *et al.*, 2014).

Dans la suite de ce manuscrit, nous décrivons les mesures les plus connues : *PMI*, *NWD*, *LSA* et *Word2vec*.

6.2.1 PMI (Pointwise Mutual Information)

La mesure *PMI* (Church et Hanks, 1989) est la façon la plus simple pour détecter automatiquement des collocations, c'est-à-dire des séquences de mots qui apparaissent le plus souvent ensemble. *PMI* calcule la distance entre deux mots x et y en divisant la probabilité de les observer conjointement $(p(x,y))$ par le produit des probabilités d'observer séparément les mots $(p(x) * p(y))$, ce qui donne l'équation 6.1

$$PMI(x,y) = \frac{p(x,y)}{p(x) * p(y)} \tag{6.1}$$

Si les deux mots x et y ne figurent jamais ensemble, *PMI* prend la valeur 0. Une valeur proche de 1 signifie que les deux mots apparaissent souvent ensemble.

6.2.2 NWD (Normalized Web Distance)

NWD (Vitányi *et al.*, 2009), nommée aussi distance de *Vitanyi*, calcule le degré d'association entre les mots d'une collection de pages web en se basant sur leur co-occurrence dans les pages web indexées par le moteur de recherche Google. La mesure *NWD* est donnée par :

$$NWD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \tag{6.2}$$

où $f(x,y)$ est le nombre de pages contenants à la fois les mots x et y ;

$f(x)$ est le nombre de pages dans lesquelles le mot x a été observé ;

N est le nombre de pages indexées.

Pour les mots indépendants, *NWD* a une valeur élevée. À l'opposé, une valeur faible signifie que les deux mots apparaissent souvent ensemble. Comme nous nous intéressons au degré d'association, nous définissons :

$$R_{NWD} = 1 - NWD(x,y) \tag{6.3}$$

La distance *NWD* a été appliquée avec succès dans divers domaines applicatifs comme la classification des requêtes web (Rose et Chandran, 2012) et la

détection d'erreurs dans les transcriptions automatiques de la parole (Fusayasu *et al.*, 2015). Dans ce travail, les auteurs utilisent la distance *NWD* et proposent une version nommée *NRD* dans laquelle les fréquences des mots de *NWD* (voir l'équation 6.2) ont été remplacées par la pondération *TF-IDF* des mots.

6.2.3 Analyse Sémantique Latente

L'analyse sémantique latente, en anglais *Latent Semantic Analysis (LSA)*, est une technique qui permet de construire une représentation sémantique d'un corpus *via* une matrice d'occurrences des mots d'un ensemble de documents. La LSA réduit l'espace de représentation d'un corpus et permet d'établir automatiquement des relations entre les termes du corpus.

Le principe général de *LSA*
Soit M la matrice de la représentation vectorielle du corpus. Les lignes représentent les mots et les colonnes sont les documents. L'élément $a_{i,j}$ de la matrice représente le nombre d'occurrences du mot i dans le document j.

$$M = \begin{array}{c} \\ w_1 \\ \\ \\ w_n \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & ... & ... & d_m \end{array} \\ \begin{pmatrix} a_{1,1} & ... & ... & a_{1,m} \\ & ... & ... & \\ & ... & ... & \\ a_{n,1} & ... & ... & a_{n,m} \end{pmatrix} \end{array}$$

La LSA effectue une décomposition en valeurs singulières de la matrice M qui donne deux matrices orthogonales U et W et une matrice diagonale Σ. On a alors :

$$M = U\Sigma V^t$$

où U est une matrice orthogonale de taille $n \times n$
Σ est une matrice diagonale de taille $n \times m$. Les valeurs des éléments diagonaux sont supérieures à 0. La matrice Σ est de taille $n \times m$
V^t désigne la transposée de la matrice orthogonale V de taille $m \times m$.
Avec $U^tU = V^tV = \mathbf{I}$ ($\mathbf{I}$ désigne la matrice d'identité).

Les éléments diagonaux de la matrice Σ contiennent les valeurs singulières qui sont ordonnées [1] de la plus grande à la plus petite valeur (c'est-à-dire que $\Sigma_{1,1}$ contient la plus grande valeur de Σ). Une approximation de la matrice M est définie en réduisant le rang de la matrice Σ aux k ($k < d$) premiers éléments diagonaux non nuls. En d'autres termes, M_k correspond aux k premières colonnes de la matrice M.

1. Pour cela, on effectue une permutation des colonnes de la matrice Σ. Les permutations sont reportées sur les colonnes de la matrice U et V.

$$M \simeq M_k = U_k \Sigma_k V_k^t$$

La matrice M_k est beaucoup moins creuse que la matrice M.
L'Analyse Sémantique Latente est principalement utilisée pour calculer la similarité entre deux phrases projetées sur l'espace des concepts, ce qui permet d'enrichir leurs représentations vectorielles.
La distance *cosinus* est utilisée pour calculer la similarité entre les paires de mots. Plus formellement, la distance entre les mots w_i et w_j est donnée par :

$$R_{LSA}(w_i, w_j) = cosine(w_i, w_j) \tag{6.4}$$

où w_l est la représentation vectorielle du mot w_l correspondant à la $l^{ème}$ ligne de la matrice M_k.

6.2.4 Word2vec

Dans (Mikolov *et al.*, 2013a), les auteurs décrivent leur outil *word2vec* [2]. Ce dernier, permet de donner aux mots une représentation vectorielle dans un espace de n dimensions appris à partir de réseaux de neurones. Plusieurs travaux récents ont montré que l'outil *word2vec* a permis d'améliorer plusieurs tâches dans le domaine du traitement automatique des langues écrites et parlées comme la traduction automatique (Kågebäck *et al.*, 2014), la reconnaissance des entités nommées (Siencnik, 2015), *etc.* Pour calculer la distance sémantique entre deux mots, *word2vec* utilise la mesure *cosinus* appliquée sur la représentation vectorielle de ces mots.

$$R_{word2vec}(x, y) = cosine(V(x), V(y)) \tag{6.5}$$

où $V(x)$ est la représentation vectorielle du mot x.
Deux mots sont considérés comme similaires si leurs représentations vectorielles sont proches dans le même espace continu.
L'outil *word2vec*, propose deux architectures : *CBOW* (Continuous Bag-Of-Words) et *skip-gram*. Chacune de ces architectures est composée de trois couches :

— La couche d'entrée prend une séquence de mots $w_1, w_2, ..., w_T$ (où T correspond à l'ensemble des mots dans le corpus). Chaque mot est représenté par un vecteur de taille V (où V est la taille du vocabulaire). Sachant que pour chaque vecteur, un seul élément prend la valeur 1 (pour indiquer la présence du mot), les autres sont nuls. L'activation des neurones dépend de l'architecture choisie.

2. http ://code.google.com/p/word2vec/

— La couche cachée projette les mots dans un espace de dimension réduite de taille N. Le nombre de neurones est un paramètre à définir dans l'outil *word2vec*.
— La couche de sortie permet de prédire soit un mot à partir de son contexte : c'est le cas pour l'architecture *CBOW*, soit l'inverse avec *skip-gram* (*i.e* prédire le contexte d'un mot).

L'architecture CBOW

Le modèle *CBOW* consiste à entraîner un réseau de neurones à prédire un mot à partir de son contexte (voir la figure 6.2). Les poids entre la couche d'entrée et la

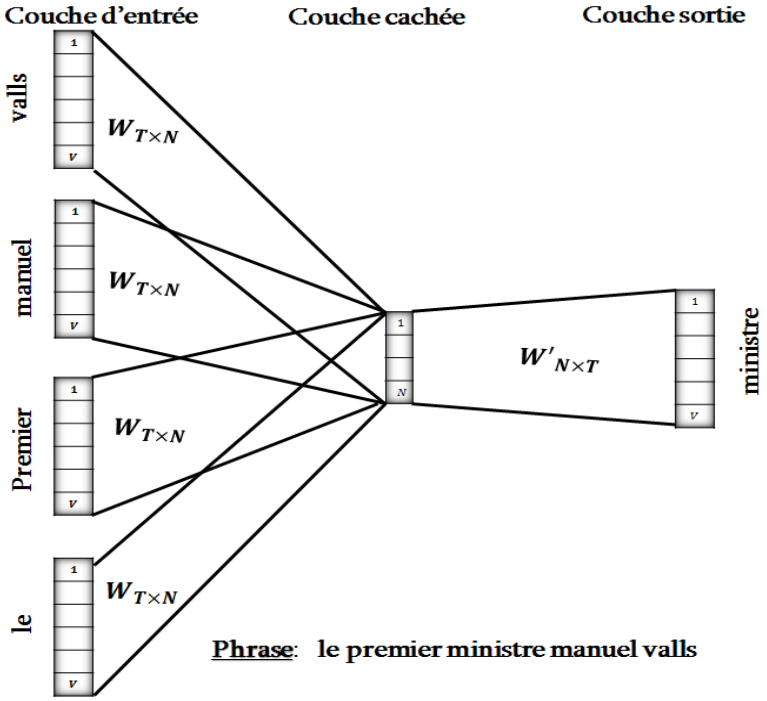


FIGURE 6.1 – *Exemple de l'architecture CBOW de Word2vec*

.

couche cachée sont représentés par la matrice W de taille $T \times N$. Les neurones de la couche cachée combinent les poids par le biais d'une fonction d'activation *linéaire* (*i.e.* la somme des poids d'entrées). Ensuite, les poids sont présentés à la couche de sortie afin de les modifier en comparant le score calculé au mot à prédire. Puis, la technique de *rétropropagation* du gradient est appliquée pour calculer le gradient de l'erreur pour chaque neurone du réseau. Cette technique permet de rétro-propager l'erreur commise de la couche de sortie vers l'arrière jusqu'à la couche d'entrée tout en modifiant les poids des neurones.

Afin de transformer la sortie du réseau en une distribution de probabilité, deux solutions ont été proposées : soit l'utilisation de la version hiérarchique de

softmax (Morin et Bengio, 2005), soit l'utilisation de l'algorithme *Negative Sampling* (Mikolov *et al.*, 2013b).
L'architecture *CBOW* cherche à maximiser la moyenne de *log* probabilité :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\ j\neq 0} log\ p(w_t|w_{t+j}) \tag{6.6}$$

où c est la taille maximale de la fenêtre de contexte.
Le paramètre c de l'outil *word2vec* permet de réduire le nombre de connexions (uniquement les neurones de contexte seront activés), ce qui permet de limiter le calcul, et par conséquent le temps d'apprentissage.

L'architecture skip-gram
Le principe du modèle *Skip-gram* est l'inverse de celui de *CBOW*. Un réseaux de neurones est entrainé pour prédire le contexte d'un mot donné (voir la figure 6.2). L'outil *Word2vec* présente à la couche d'entrée du modèle *skip-gram* le mot et donne en sortie $2 * c$ mots où c est la taille maximale de la fenêtre de contexte. Plus formellement, pour un corpus constitué d'une séquence de mots $w_1, w_2, ..., w_T$, le modèle *Skip-gram* maximise l'équation 6.7 :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\ j\neq 0} log\ p(w_{t+j}|w_t) \tag{6.7}$$

6.2.5 Discussion

Nous venons de voir que la distance *NWD* se base sur les collocations des mots et que *Word2vec* repose sur la représentation vectorielle de chaque mot (on parle de *word embedding*). La distance *NWD* considère les mots apparaissant souvent ensemble comme proches sémantiquement, alors que, *Word2vec* considère que plus les mots partagent des contextes similaires, plus ils sont liés sémantiquement. Prenons les deux documents suivants :

— d1 : La visite d'état du président François Hollande en Algérie.
— d2 : Le président Barack Obama a attaqué de front Donald Trump.

Avec une fenêtre de contexte de taille 5 (2 mots à gauche et 2 à droite). Les mots "visite état" et "François Hollande" sont les contextes du mot président pour le document $d1$ et "Barack Obama" pour le document $d2$ (contexte gauche est vide). Puisque les mots *Barack* et *Obama* apparaissent dans le même contexte que les mots *François* et *Hollande*, l'approche *Word2vec* considère que ces mots

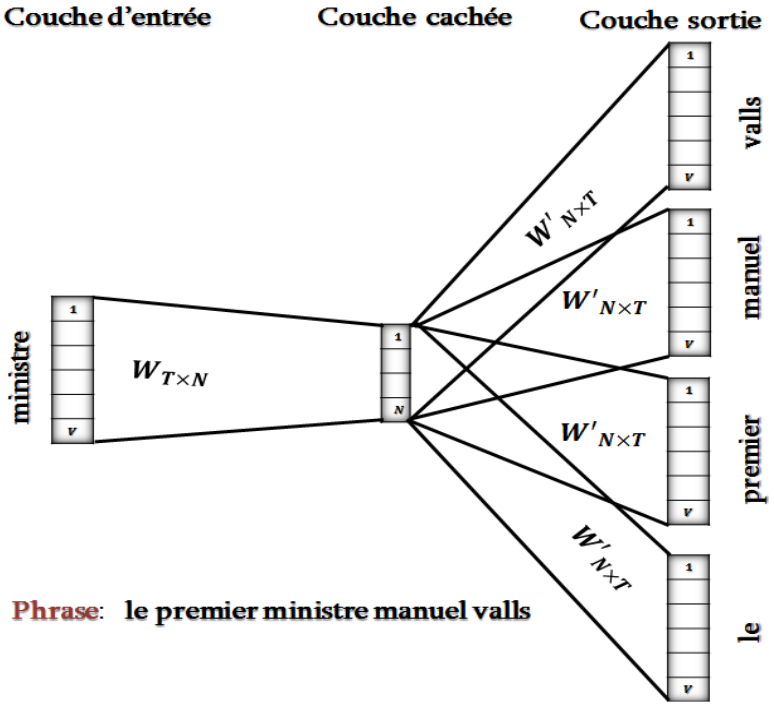


FIGURE 6.2 – *Exemple de l'architectures et Skip-gram de Word2vec*

.

sont similaires. Alors que, dans cet exemple la distance *NWD* n'établit aucun lien sémantique entre les mots. En ce qui concerne la segmentation thématique, les relations sémantiques de type *contexte* peuvent produire du bruit. En effet, la similarité entre deux parties de l'émission qui traitent de deux thèmes différents peut prendre de valeurs fortes avec des relations de contexte. Le système de segmentation thématique peut alors croire que les deux parties traitent du même thème.
Afin de profiter des points forts de chaque approche, nous effectuons des pré-traitements permettant de ne garder que les mots pleins du corpus. Pour cela, nous supprimons les mots outils et nous remplaçons les mots par leur lemme (voir la sous-section 1.3.1). Cela permet d'avoir une notion de contexte moins liée à la syntaxe pour *Word2vec*. Nous proposons aussi de remplacer la structuration du corpus en phrase par une structuration par article et d'utiliser un contexte de mots un peu élargi. Ces pré-traitements peuvent rendre la représentation vectorielle beaucoup plus robuste et appropriée à la détection des changements de thèmes. En appliquant l'ensemble des pré-traitements, l'outil *Word2vec* aura la possibilité de capter plus de mots ayant un lien thématique avec le mot à prédire.

6.3 Méthodes de segmentation thématique basées sur la distance entre les mots

Cette famille de méthodes s'intéresse non seulement à la répétition des mots s'écrivant de la même façon mais s'étend également à d'autres types de relations comme la synonymie. Pour cela, ces méthodes utilisent d'autres ressources linguistiques comme des dictionnaires ainsi que des corpus généraux ou spécifiques. Dans l'état de l'art, il existe plusieurs algorithmes parmi lesquels on peut citer (Ferret, 2002), (Stokes, 2003), (Guinaudeau *et al.*, 2012) et (Sakahara *et al.*, 2014). Dans cette section, nous donnons un petit panorama de ce genre d'algorithmes.

L'auteur de (Stokes, 2003) établit les chaînes lexicales non seulement à partir des répétitions de mots mais également avec les mots partageant des liens sémantiques. Ces relations sont obtenues manuellement à partir de *WordNet*. Dans d'autres travaux, les relations ont été acquises de façon automatique par le biais d'articles de presse ou de contenus *Wikipédia*.

Dans (Choi *et al.*, 2001), les auteurs intègrent dans leur algorithme *C99* (voir la sous section 1.4.1) les relations sémantiques entre les mots par l'utilisation de l'analyse sémantique latente. Pour rappel, la technique *LSA* construit pour chaque mot un vecteur représentant sa distribution dans les contextes, les mots apparaissant dans le même contexte sont considérés comme similaires. En d'autres termes, ces vecteurs ayant une représentation vectorielle similaire ou proche sont considérés comme liés sémantiquement. Les auteurs utilisent un corpus constitué de $104k$ phrases correspondant à $35k$ paragraphes pour effectuer la décomposition en valeurs singulières de la matrice initiale M (mots-documents). Cette opération donne une matrice M_k de dimension réduite k. Soit $f_{i,j}$ la fréquence du terme j dans la représentation vectorielle de la phrase s_i. L'intégration des relations estimées par *LSA via* la matrice M_k est donnée par :

$$\lambda_i = \sum_j f_{i,j} \times M_k(j) \tag{6.8}$$

La cohésion lexicale entre deux phrases s_i et s_j représentées respectivement par les vecteurs λ_i et λ_j est calculée à partir de la mesure *cosinus*.

Le système de segmentation thématique développé dans (Guinaudeau *et al.*, 2012) repose sur l'algorithme de *Utiyama* (Utiyama et Isahara, 2001) présenté dans 1.4.1. Pour avoir un système beaucoup plus performant, (Guinaudeau *et al.*, 2012) a proposé plusieurs améliorations parmi lesquelles l'intégration de relations sémantiques. Les auteurs proposent de modifier les comptes des mots lors de l'estimation du modèle de langue, ce dernier est utilisé par la suite dans le calcul de la cohésion lexicale. Soit w_i^k le compte du $k^{ème}$ mot du segment S_i

(*i.e* le nombre de fois où le mot apparaît dans un segment). L'intégration de relations pour le mot w_i^k est donnée par :

$$(w_i^k)' = w_i^k + \sum_{j=1, w_i^j \neq w_i^k}^{n_i} r(w_i^j, w_i^k) \tag{6.9}$$

où n_i est le nombre de mots dans le segment S_i
$r(w_i^j, w_i^k) \in [0;1]$ est la valeur de similarité entre les mots w_i^j et w_i^k.
Les distances entre les mots ont été estimées à partir des articles du journal *Le Monde* et la transcription manuelle des campagnes *ESTER1* et *ESTER2*.
Dans (Ferret, 2002), les relations ont été utilisées à l'aide d'une fenêtre glissante pour comparer le contexte thématique du vecteur du texte (les mots figurant dans la fenêtre) et celui du segment (constitué à partir des mots du réseau de collocations). Les relations sont calculées à partir des articles *Le Monde* (entre 1990 et 1994) pour le français et les articles du *L.A. Times* pour l'anglais.

Récemment, (Sakahara *et al.*, 2014) ont proposé un algorithme de segmentation qui exploite les relations entre les mots pour construire des clusters thématiques pour chaque phrase du document à segmenter. Les auteurs considèrent que deux phrases sont thématiquement homogènes si et seulement s'il y a une forte corrélation entre leurs clusters thématiques. Les distances entre les clusters sont estimées à partir des similarités entre les mots. Deux types de relations entre les mots ont été prises en compte : la similarité sémantique obtenue à partir de *Word2vec* (voir la section 6.2.4) avec une fenêtre de 5 mots, et la similarité de cooccurrence qui mesure le nombre de fois où les mots figurent ensemble, l'un à côté de l'autre avec une fenêtre glissante centrée de taille 5.
(Nie *et al.*, 2013a) utilisent leur propre algorithme (que nous ne détaillons pas) pour intégrer les relations sémantiques. La matrice de relations est intégrée au moment même du calcul de similarité. Plus formellement, la similarité entre deux vecteurs s_i et s_j, avec la prise en compte de la matrice de relations M est donnée par :

$$sim(s_i, s_j|S) = \frac{s_i^T \; M \; sj}{\|s_i\| \|s_j\|} \tag{6.10}$$

Plusieurs travaux comme (Ferret, 2002) et (Misra *et al.*, 2009) montrent que dans la tâche de segmentation thématique, les ressources générales n'améliorent souvent pas la performance du système.

Dans ce chapitre, nous utilisons des relations extraites le même jour que le document à segmenter. Ce mécanisme présente un double avantage : d'une part, il permet de limiter le nombre de relations hors domaine. D'autre part, il permet d'avoir un nombre très important de relations pertinentes à jour.

6.4 Calcul de similarité sémantique

Notre système tel que nous l'avons présenté jusqu'à maintenant est basé sur le calcul de la cohésion de la parole entre deux blocs tout au long de l'émission. Pour rappel, la similarité entre deux blocs b_i et b_j est donnée par :

$$Sim(b_i, b_j) = \frac{\nu(b_i^T) \times \nu(b_j)}{\sqrt{\nu(b_i^T) \times \nu(b_i)} \times \sqrt{\nu(b_j^T) \times \nu(b_j)}} \tag{6.11}$$

où : $\nu(b_i)$ est la représentation vectorielle du bloc b_i et b_i^T son transposé. b_i est de taille $n \times 1$ (n est la taille de vocabulaire).
Si le mot m apparaît au moins une fois dans le bloc b_i, $\nu(b_{i(m)})$ prend une valeur dans l'intervalle $]0, 1]$, sinon il vaut zéro.

Comme il a été déjà mentionné précédemment, la cohésion lexicale telle qu'elle est définie dans l'équation 6.11 ne prend pas en compte les relations entre les mots. Dans (Nie *et al.*, 2013b), les auteurs intégrent les relations directement dans le calcul de la cohésion entre les phrases selon l'équation suivante :

$$Sim(b_i, b_j|M) = \frac{\nu(b_i^T) \times M \times \nu(b_j)}{\sqrt{\nu(b_i^T) \times \nu(b_i)} \times \sqrt{\nu(b_j^T) \times \nu(b_j)}} \tag{6.12}$$

où M est la matrice de relations entre les mots.

Nous proposons une extension de cette approche en intégrant dans l'équation 6.12 la matrice de relations au niveau du dénominateur pour les deux blocs, ce qui donne l'équation 6.13 :

$$Sim(b_i, b_j|M) = \frac{\nu(b_i^T) \times M \times \nu(b_j)}{\sqrt{\nu(b_i^T) \times M \times \nu(b_i)} \times \sqrt{\nu(b_j^T) \times M \times \nu(b_j)}} \tag{6.13}$$

Notre proposition présente comme avantage d'assurer l'intégration de la similarité sémantique à deux niveaux :

- *Intra-bloc* : la projection dans l'espace sémantique ne concerne que les mots du même bloc, c'est-à-dire que l'intégration de relations se passe au niveau de chaque bloc séparément. La prise en compte des relations sémantiques est assurée par le produit vectoriel du bloc représenté par un vecteur avec son transposé $\nu(b_i^T) \times M \times \nu(b_i)$ et $\nu(b_j^T) \times M \times \nu(b_j)$ respectivement pour les blocs b_i et b_j.
- *Extra-bloc* : la projection dans l'espace sémantique ne concerne que les mots ayant des relations avec les mots de l'autre bloc. Ceci permet de renforcer des liens existant entre les deux blocs b_i et b_j. L'intégration est assurée par le produit vectoriel $\nu(b_i^T) \times M \times \nu(b_j)$.

6.5 Corpus d'apprentissage

Comme nous avons montré dans la section 6.3, la plupart des travaux exploitent des relations provenant de données générales (articles de presse, Wikipédia, *etc.*). Malgré l'importance du modèle de langage utilisé dans l'extraction de relations, aucun travail à notre connaissance n'a mis l'accent sur son adaptation au type de documents à segmenter.
En ce qui concerne les journaux d'information télévisés, les relations entre les mots sont généralement extraites à partir de ressources (articles de presse, transcription manuelle des émissions radiophonique) d'une précédente période par rapport à la date de diffusion du journal télévisé. Cette différence peut aller jusqu'à 10 ans comme, les travaux présentés dans (Ferret, 2002). Cependant, les journaux télévisés se focalisent beaucoup plus sur les thèmes d'actualités. Par nature, ce type d'informations est dynamique dans le temps et même dynamique dans une journée au fur et à mesure de l'avancement des événements. Par exemple, la relation entre les mots : *(Malaysia, boîte)*, *(Malaysia, noir) (MH370, noir)*, *etc.* est plus forte après le crash puis après le lancement de la recherche de la boîte noire. D'autres mots comme *France* et *médaille* peuvent être liés dans un corpus contenant des données générales, mais en période de jeux olympiques, la relation entre les deux mots se renforce. Cela revient essentiellement au nombre très important d'articles traitant de la participation de l'équipe de France à ce moment précis.
Dans nos expériences, nous avons utilisé deux corpus pour calculer la matrice de similarité. Le premier nommé *GenNews* est de grande taille et est constitué de plusieurs sources d'informations (articles de presses et transcriptions manuelles de journaux télévisés) liées à l'actualité. Le deuxième, nommé *DiaNews*, est plus petit et est collecté à partir des articles de *Google Actualités* du même jour que l'émission à segmenter.

6.5.1 Données d'actualité générales (*GenNews*)

Le corpus *GenNews* est extrait à partir de différentes sources :
- Articles de presse « Le Monde » édités entre 1987 et 2003.
- Articles de presse de « Google Actualités » publiés entre 2010 et 2012.
- 400 heures de transcription manuelle de journaux d'informations datant de 2005 à 2012 et provenant des campagnes d'évaluation *ESTER1*, *ESTER2*, *EPAC* et *ETAPE*.

Au total, le corpus contient environ 2 milliards de mots distribués en 66 millions de lignes. Après les pré-traitements (suppression des mots vides et lemmatisation), nous obtenons 1069 millions d'occurrences, sachant que la taille du

vocabulaire est de l'ordre 3,73 millions de mots.

6.5.2 Données diachroniques (*DiaNews*)

Les données diachroniques (*DiaNews*) ont été largement utilisées dans la transcription automatique de journaux télévisés ((Allauzen et Gauvain, 2005), (Illina *et al.*, 2014)) pour mettre à jour le modèle de langage et limiter le nombre de Mots Hors Vocabulaire. Nous pensons que ce type de documents peut également présenter un avantage dans le cadre de la segmentation thématique.

Durant la semaine du 10 au 16 février et avec un rythme d'une fois par heure, nous avons collecté, à travers le site web multi-domaines *Google Actualités* une base de données de 39350 articles, soit en moyenne 5620 par jour. Un article de presse est alors représenté par un tuple dans une base de données celui-ci contient l'identifiant de l'article, le lien de la page web, le titre de l'article, sa date de publication et l'identifiant de l'article principal, qui est issu d'un processus de regroupement d'articles. En effet, au delà de la collecte des articles d'information en provenance du web, *Google Actualités* regroupe les articles traitant du même thème. Chaque cluster est représenté par un seul article considéré comme le principal (voir l'exemple illustré dans la figure 6.3).

Les pages de la presse en ligne étant bruitées par des éléments non-informatifs, l'extraction du contenu de l'article à partir des données brutes est une nécessité pour les traitements ultérieurs. Pour cela, certains outils sont disponibles : *Boilerpipe* [3], *JusText* [4], *Readability* [5], *Diffbot* [6]. Dans nos travaux, nous avons choisi d'utiliser l'outil *Boilerpipe* a été utilisé pour déterminer la pertinence d'un bloc textuel à partir d'un arbre de décision.

Le tableau 6.1 donne des statistiques sur la nature du corpus *DiaNews* après les pré-traitements (suppression des mots fréquents et lemmatisation). En pra-

	# articles	# mots	# mots différents
Moy. par jour	5.6*K*	692*K*	11*K*
Min	4.0*K*	482*K*	9.1*K*
Max	6.5*K*	800*K*	12.3*K*

Tableau 6.1 *– Statistiques sur le corpus DiaNews.*

tique, pour chaque émission, nous sélectionnons un sous-ensemble de relations

3. https ://code.google.com/p/boilerpipe/
4. https ://code.google.com/p/justext/
5. https ://code.google.com/p/arc90labs-readability/
6. https ://pypi.python.org/pypi/diffbot

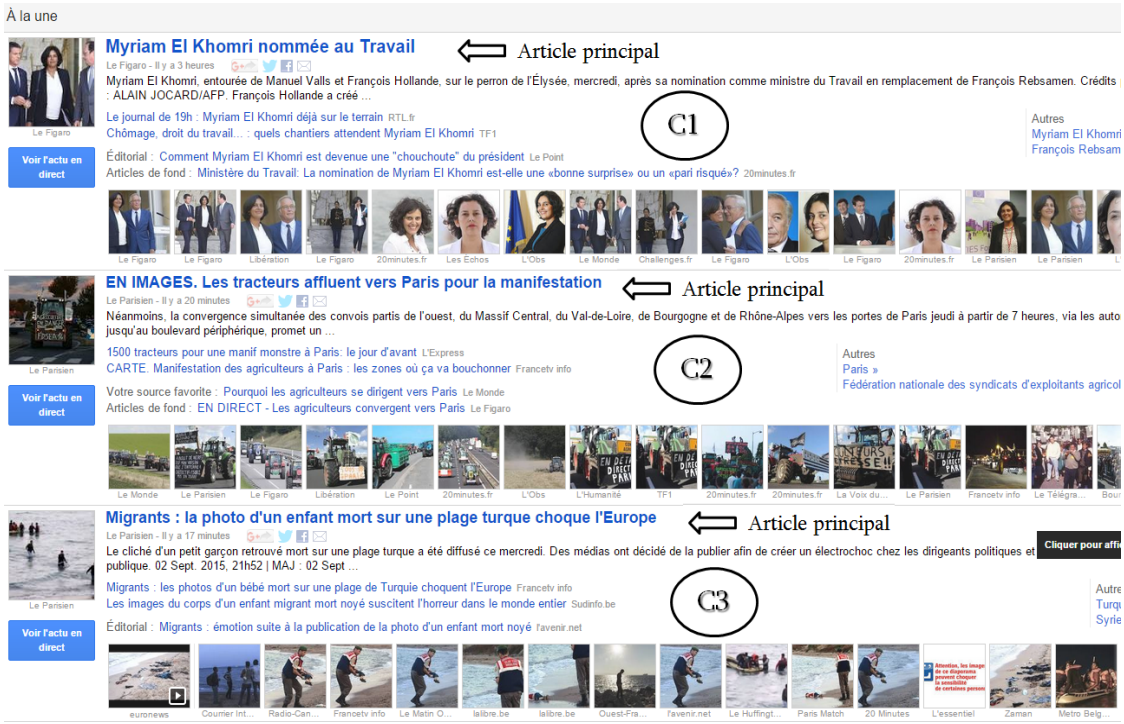


FIGURE 6.3 – *Clustering des articles de presse par Google Actualités.*

correspondant aux mots prononcés dans l'émission. En effet, le fichier de relations obtenu après le calcul des distances entre les mots est très volumineux. Par conséquent, notre algorithme met beaucoup de temps pour chercher les liens sémantiques *utiles*. Une relation sémantique est considérée comme utile si les deux mots figurent dans l'émission à segmenter. C'est la raison pour laquelle nous sélectionnons un sous-ensemble de relations ce qui permet de réduire le temps d'exécution.

6.5.3 Comparaison des relations sémantiques entre les mots selon les corpus

Le tableau 6.2 illustre quelques relations issues des données *DiaNews* et *GenNews* enrichissant une émission du corpus MCS7-14 (le journal d'*Arte* du *10/02/2014*). Parmi les thèmes traités dans l'émission, on trouve la visite du président François Hollande aux États-Unis, les négociations sur la Syrie qui se sont déroulées à Genève et la célébration de la révolution Iranienne. Nous constatons que les relations extraites à partir des données *diachroniques* sont plus pertinentes et adaptées aux thèmes évoqués dans l'émission par rapport aux données générales. Par exemple, la visite du président François Hollande aux USA fait

Données diachroniques - DiaNews					
genève	syrien	0.55	hollande	visite	0.60
genève	bachar	0.54	hollande	barack	0.54
genève	assad	0.52	hollande	françois	0.51
genève	négociation	0.51	hollande	voyage	0.43
genève	lakhdar	0.50	françois	obama	0.48
genève	accord	0.49	françois	président	0.48
genève	contingent	0.49	françois	barack	0.48
genève	brahimi	0.48	françois	visite	0.44
genève	damas	0.48	président	françois	0.48
genève	délégation	0.48	président	chef	0.46
genève	discussion	0.43	président	barack	0.46
genève	transition	0.42	président	bachar	0.44
genève	syrie	0.42	téhéran	iran	0.65
genève	rohani	0.41	téhéran	rohani	0.65
genève	régime	0.41	téhéran	occidentaux	0.64
genève	médiateur	0.41	téhéran	balistique	0.59
genève	nucléaire	0.39	téhéran	iranien	0.57
genève	iran	0.39	téhéran	hassan	0.57
genève	hassan	0.39	téhéran	nucléaire	0.54
Données générales - GenNews					
genève	bâle	0.63	hollande	moscovici	0.53
genève	berne	0.58	hollande	candidat	0.39
genève	bruxelles	0.52	hollande	droite	0.35
genève	onu	0.51	françois	présidence	0.16
genève	suisse	0.46	françois	guillaume	0.51
genève	international	0.44	président	présidence	0.81
genève	représentant	0.41	président	ministre	0.63
genève	délégation	0.39	téhéran	iran	0.85
genève	bilatéral	0.37	téhéran	iranien	0.84
genève	téhéran	0.36	téhéran	iranien	0.84

Tableau 6.2 – *Exemples de similarité entre les mots extraits à partir des données GenNews et DiaNews.*

apparaître de nouvelles relations comme (hollande, visite), (hollande, voyage), *etc.* À l'inverse, les données générales contiennent des relations inadaptées au contexte du segment en question et couvrent moins le contenu de l'émission à segmenter. Nous arrivons au constat pour le thème, qui concerne les négociations syriennes en Suisse.

6.6 Résultats et discussion

6.6.1 Cadre expérimental

Dans cette étude, nous comparons l'apport des relations sémantiques provenant de *NWD* et *Word2vec*. Nous choisissons l'architecture *CBOW* qui a pour avantage d'être rapide par rapport à *skip-gram*. Nous fixons la taille maximale de la fenêtre du contexte à 17 mots et nous utilisons la technique d'échantillonnage négatif pour apprendre le modèle. Comme les corpus utilisés pour les expérimentations sont petits, nous limitons la taille des vecteurs à 100.

Les expériences sont menées sur le corpus décrit dans 2.2. Lors de la collecte du corpus *GenNews*, la notion de document a été perdue. Nous ne sommes donc pas en mesure d'estimer les distances *NWD* de ce corpus. En revanche, plusieurs configurations ont été évaluées avec la distance *Word2vec* :

- *W2V_GenNews* entraîné à partir du corpus *GenNews*.
- *W2V_DiaNews* appris à partir du corpus *DiaNews* pour lequel les articles de presse du même jour sont concaténés dans un seul fichier.
- *W2V_DiaNews_1* est une variante de *W2V_DiaNews* à ceci près qu'elle supprime la structuration en phrases (*i.e* tous les mots d'un article donné ont été placés sur une seule ligne). En d'autres terme, le corpus d'apprentissage est structuré en articles et non pas en phrases.

L'intérêt de placer les lemmes retenus d'un article dans une seule ligne (*i.e.* la configuration *W2V_DiaNews_1*) est de capturer plus de mots ayant un lien thématique avec le mot à prédire, même s'ils sont situés dans des phrases adjacentes.

6.6.2 Résultats et discussion

La figure 6.4 donne les performances du système en terme de *Rappel/Précision* au niveau des frontières. Les courbes ont été obtenues en faisant varier la valeur du seuil de segmentation. La configuration *Without* correspond au système de segmentation sans matrice de relations. L'utilisation des relations issues des

données d'actualité générales (*GenNews*) améliore la *Précision*, ce qui permet d'avoir une meilleure détection de ruptures. En revanche, ce type de relations a tendance à dégrader le *Rappel*. Ceci est principalement dû au fait que les relations induites par ce corpus sont trop bruitées. Elles ont donc tendance à associer trop de mots dans des thèmes différents. En ce qui concerne l'utilisation du

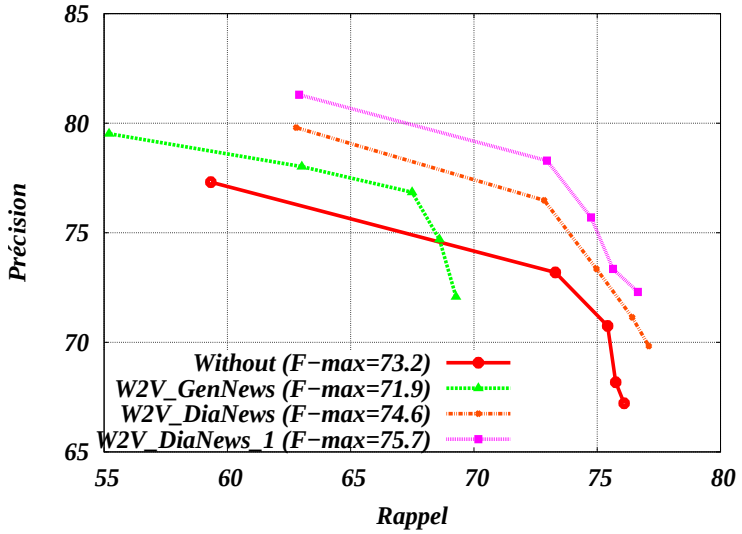


FIGURE 6.4 – *Performance du système avec les différentes configurations de Word2vec.*

corpus *DiaNews*, les deux configurations de *Word2vec* améliorent la qualité de la segmentation ; la performance la meilleure est réalisée avec la configuration *W2V_DiaNews_1*.

Nous pouvons rapidement observer que les relations issues du *DiaNews* augmentent principalement la *Précision* sans trop dégrader le *Rappel*. Cela peut être expliqué par le fait que le système basé uniquement sur la redondance des mots et de locuteurs (*without*) produit un nombre important de fausses alarmes par rapport au système de segmentation qui prend en compte les relations sémantiques. En effet, à défaut de répétitions au niveau des mots et des locuteurs, les valeurs de la cohésion prennent des valeurs faibles. Ceci conduit le système à annoncer des changements de thèmes alors que dans la référence il n'y a aucune rupture. En revanche, les relations sémantiques augmentent les valeurs de la cohésion ; ainsi le système considère que les blocs en question traitent de la même thématique.

La figure 6.5 représente la courbe de la cohésion de la parole sans (a) et avec (b) relations sémantiques de l'émission TF1_Journal13Heures. Les lignes verticales de la figure correspondent aux frontières de référence et les ovales représentent les frontières d'hypothèse. Nous constatons que :

— Dans la zone (1), il y a les thèmes suivants :

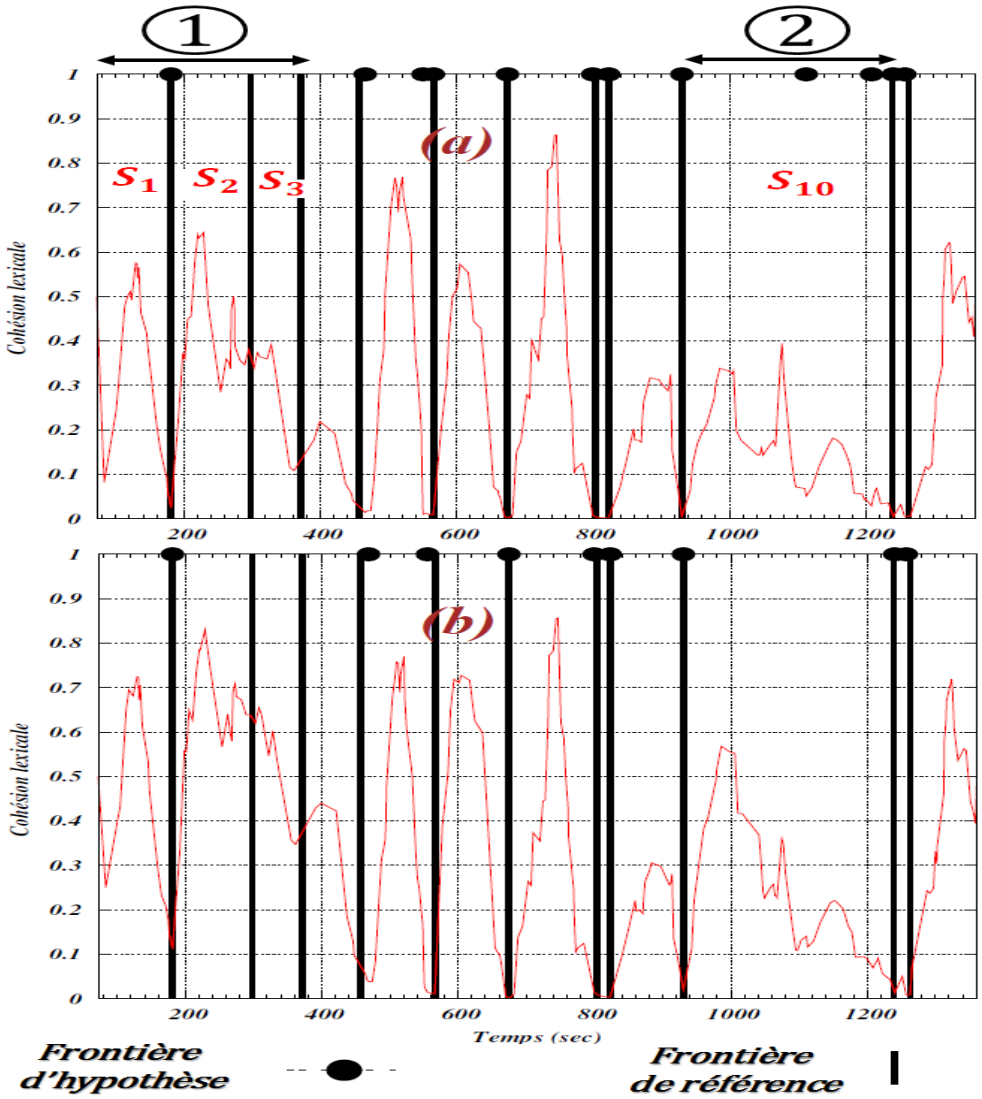


FIGURE 6.5 – *Cohésion de parole sans relations sémantiques (a), et avec relations sémantiques (b) de l'émission TF1_Journal13Heures_2014-02-15. Les relations sémantiques ont été calculées à partir de la configuration w2v_DiaNews.*

S1. Tempête Ulla en Bretagne,

S2. Conséquences de la tempête Ulla à Rennes, Brest sur le trafic ferroviaire,

S3. Décès en mer cette nuit à cause de la tempête.

— La zone (2) porte sur « le salon de l'agriculture ».

Les segments consécutifs qui traitent de thèmes très proches thématiquement sont difficiles à identifier (voir la zone (1)) que ce soit avec ou sans relations sémantiques. En effet, comme les segments partagent plusieurs mots, la

détection de frontières devient de plus en plus difficile (*i.e* le système met en évidence peu de changements de thèmes). Par exemple, le thème portant sur les intempéries en Bretagne (Ouest de la France) et les intempéries en Grande-Bretagne peuvent avoir plusieurs mots en commun (*dégâts, inondation, endommager, etc.*). L'ajout des liens thématiques entre les mots augmente la cohésion entre les segments et provoque une erreur de type *suppression*.
La zone (2) illustre l'effet positif des relations sémantiques. Le système basé uniquement sur la réitération de mots ayant la même forme écrite produit des fausses frontières. En revanche, avec le renforcement de la cohésion par des relations sémantiques, le système n'a commis aucune erreur.

À noter que le gain apporté par les relations sémantiques pourrait être beaucoup plus important si les particularités suivantes étaient moins fréquentes :

— Le nombre important de thèmes non traités dans les articles de presse. Dans ce cas, la cohésion prend des valeurs quasi similaires avec ou sans relations sémantiques.

— Le système de segmentation ne profite pas pleinement de toutes les relations extraites à partir des articles de presse. Cela revient essentiellement aux erreurs de transcription automatique. Le nom *Maria Riesch* est ainsi transcrit *Maria Riche* ce qui a pour conséquence que toutes les relations qui ont un lien avec *Riesch* ne sont pas prises en compte.

Il est intéressant de mentionner que la cohésion est calculée à partir de la distribution conjointe des mots et des locuteurs (voir le chapitre 5). Cette combinaison cache l'apport net des relations sémantiques, dans les deux cas suivants :

- 1^{er} cas : l'intervention d'un locuteur dans deux thèmes consécutifs augmente les valeurs de cohésion de la parole. Par conséquent, la détection de la frontière n'est pas certaine même si la cohésion est renforcée par des relations sémantiques.

- $2^{ème}$ cas : l'intervention d'un locuteur dans un seul thème augmente les valeurs de cohésion de la parole tout en permettant d'éviter des fausses alarmes même s'il y a très peu de répétitions et/ou de relations sémantiques.

Nous observons également que la qualité de la segmentation avec la configuration *W2V_DiaNews_1* est meilleure que celle obtenue avec *W2V_DiaNews*. Cela met en évidence que les relations acquises avec la configuration *W2V_DiaNews_1* sont meilleures. En effet, la structuration du corpus d'apprentissage en articles (*i.e.* tous les mots d'un article sont placés sur une seule ligne) aide l'outil *Word2vec* à capturer plus de régularités sémantiques sur le mot à prédire. Ceci permet de rattraper ou de renforcer la présence des mots ayant un lien sémantique avec le mot cible.

D'autres séries d'expériences ont été menées, cette fois-ci en comparant : *(i)*

la meilleure configuration obtenue de *Word2vec* avec *(ii)* les relations calculées avec la distance *NWD* et d'autres venants de *(iii)* la combinaison des deux distances (*Word2vec* et NWD). De la figure 6.6, nous constatons que *NWD* dépasse

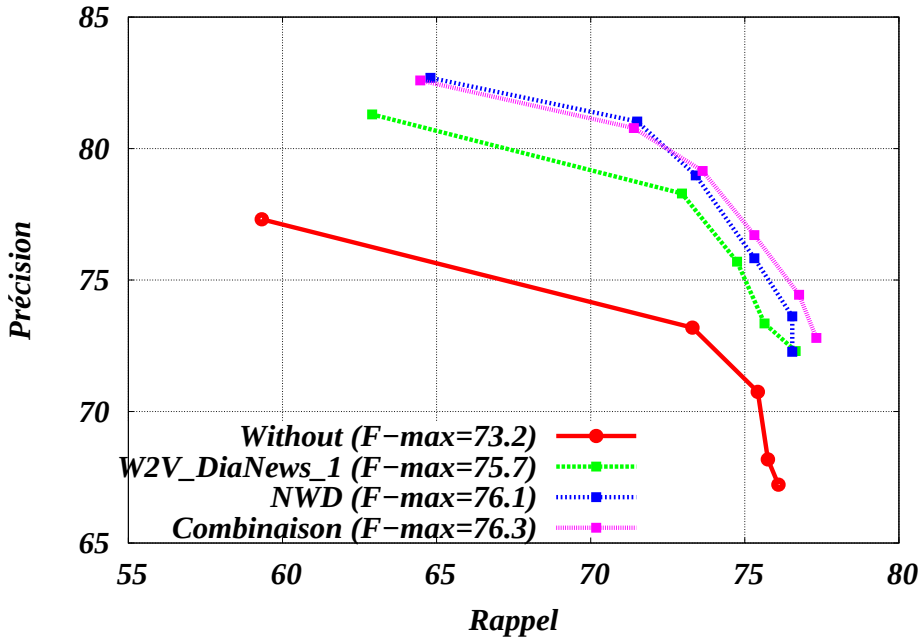


FIGURE 6.6 – *Performance du système de segmentation avec des relations calculées à partir de Word2vec et NWD.*

légèrement *Word2vec*. Cela signifie que la prise en compte des co-occurrences entre les mots (*NWD*) semble plus bénéfique que l'utilisation du contexte des mots avec l'approche neuronale. Cependant, une simple combinaison linéaire de ces deux matrices de relations (avec une interpolation linéaire de 0.5) apporte une légère amélioration.

L'évaluation en terme de *CouvN* et *CouvD* est donnée dans le tableau 6.3. Pour cela, pour chaque configuration nous avons choisi le point de fonctionnement qui maximise la *F-mesure* au niveau des frontières. L'intégration des données *DiaNews* augmente *CouvN* de 61.1% à 66.3% et *CouvD* de 70.3% à 75.1%. En revanche, les relations provenant de données d'actualité générales perturbent le système en détectant moins de frontières. Ceci entraine une dégradation de F-mesure en terme de nombre de frontières et de durée des segments correctement retournés *CouvD* avec un écart de 2.5 et 1.7 respectivement. Cependant, dans l'évaluation au niveau du nombre de segments, une légère amélioration a été constatée. Cela est principalement dû à la nature de la mesure *CouvN*, qui privilégie dans certains cas les segmentations ayant moins de segments hypothèses.

Nous avons testé notre algorithme avec les émissions du corpus *MCS5-15*

	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvD*
Without	72.7	71.8	72.2	62.0	60.3	61.1	70.3	70.3	70.3
GenNews	63.0	78.0	69.7	58.4	64.6	61.3	68.6	68.7	68.6
DiaNews	73.6	79.1	76.3	64.6	68.1	66.3	75.1	75.1	75.1

Tableau 6.3 – *Performance du système avec les trois métriques d'évaluation F-mesure, CouvN et CouvD. Les relations sémantiques utilisées sont issues de la combinaison linéaire de Word2vec et NWD pour les données DiaNews.*

sur les relations *DiaNews*. Les résultats sont donnés dans le tableau 6.4. Nous constatons que le système réagit de la même façon qu'avec le corpus *MCS7-14*. Les relations *DiaNews* améliorent la performance du système et la segmentation qui prend en compte les relations a une meilleure *Précision*.

	Nbre frontières			Nbre segments			Durée segments		
	R	*P*	*F-mesure*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CouvN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{N}$	*CouvD*
Without	68.3	63.7	65.9	58.9	55.0	56.9	63.4	63.3	63.3
DiaNews	69.0	72.9	70.9	60.0	66.5	64.7	70.6	70.8	70.7

Tableau 6.4 – *Performance du système sans et avec relations sémantiques dans le corpus MCS5-15.*

Enfin, nous évaluons le gain apporté par les relations sémantiques selon la nature des segments (courts ou longs). Du tableau 6.5, nous constatons que les relations sémantiques améliorent non seulement les segments longs mais aussi les courts.

	Corpus MCS7-14				*Corpus MCS5-15*			
	Seg. Longs		Seg. Courts		Seg. Longs		Seg. Courts	
	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$
Without	75.6	75.4	27.3	25.0	69.6	62.7	24.3	25.7
DiaNews	76.1	77.1	27.5	33.3	72.7	71.7	31.4	43.1

Tableau 6.5 – *Influence de l'intégration des relations sémantiques selon la taille des segments.*

Les relations sémantiques de type *DiaNews* permettent d'améliorer substantiellement la segmentation obtenue en retournant moins de segments qu'un système basé uniquement sur la cohésion de la parole. De plus, l'intégration de relations sémantiques permet d'augmenter le nombre de segments corrects. Par exemple, sur l'ensemble des 227 segments longs du corpus *MCS5-15*, le système basé uniquement sur la cohésion de la parole retourne 252 fragments théma-

tiques, dont 158 corrects. Le système exploitant les relations *DiaNews* trouve, quand à lui, 230 segments dont 165 sont corrects.

6.7 Conclusion

Dans ce chapitre, nous avons exploité les relations sémantiques entre les mots pour la tâche de la segmentation thématique des journaux d'informations. Le but est de renforcer encore plus le calcul de la cohésion lexicale en intégrant cette fois-ci des liens sémantiques entre les mots de l'émission. Pour cela, nous suggérons l'utilisation de données issues du corpus DiaNews provenant de la presse écrite du même jour que l'émission à segmenter. Nous avons proposé une nouvelle façon d'intégrer les relations qui assure l'intégration des liens sémantiques localement (dans les deux blocs séparément) et globalement (les liens existants entre les deux blocs). La méthode proposée intègre directement les relations lors du calcul de la cohésion lexicale. Nous avons utilisé *Word2vec* (basé sur un modèle vectoriel) et *NWD* (basé sur la Théorie de l'Information) dans le calcul du degré d'association entre les mots.

Les expériences menées sur les deux corpus *MCS7-14* et *MCS5-15* montrent l'importance du choix des données dans lesquelles on fait l'extraction des relations sémantiques. En effet, l'exploitation des liens sémantiques issus des relations diachroniques améliorent la performance du système. En revanche, les données générales génèrent des relations non adaptées à la nature de nos émissions ce qui conduit à dégrader la qualité de la segmentation. Les résultats obtenus dans le corpus de test (*MCS5-15*) valident l'importance des relations sémantiques.

Troisième partie

Structuration thématique : au-delà de la détection de frontières

Chapitre 7

Identification thématique des segments : le titrage

Sommaire

7.1 Introduction

La plupart des systèmes de segmentation thématique identifient les thèmes trouvés par des labels anonymes (thème1, thème2, ...). Donner un titre à chaque

segment de l'émission permet à l'utilisateur de prendre plus rapidement connaissance de l'intégralité des thèmes évoqués durant le journal d'information. Par conséquent, l'utilisateur peut aller directement aux thèmes qui l'intéressent.
Pour rappel, notre système de structuration thématique est composé de deux modules complémentaires : segmentation thématique et titrage. C'est ce dernier qui est chargé de donner une appellation aux segments thématiques retournés.

Dans le domaine journalistique, le titre joue un rôle primordial, car c'est lui qui donne au lecteur l'envie de lire ou non le contenu de l'article. Pour cela, le titre d'un segment thématique ou d'un article doit être correct et donner l'essentiel de l'information. En plus de cela, le titre doit piquer la curiosité de l'utilisateur ou du lecteur.

La segmentation thématique et le titrage sont largement utilisés dans les sites de chaînes d'information (rubrique *Replay*). Par exemple, la chaîne TF1 donne à l'utilisateur la possibilité de voir l'intégralité du journal télévisé ou bien chaque thème séparément. Son contenu est décrit sous forme d'une table des matières. Cette structure donne non seulement un aperçu des thèmes évoqués durant le JT, mais aussi la possibilité de voir un thème précis (voir la figure 7.1). Il faut noter que la structuration en thèmes proposée par les chaînes est entièrement manuelle.

Un système de structuration en thèmes est indispensable pour les organismes chargés d'archiver les productions audiovisuelles. En effet, face à l'explosion du volume des émissions disponibles, seuls les moyens automatiques peuvent permettre de satisfaire les besoins.

Le titrage automatique peut être aussi utilisé pour les e-mails ou les discussions de type *chat*. Ajouter un titre aux e-mails qui n'ont pas d'objet permet à l'expéditeur d'estimer son degré d'importance par rapport aux e-mails restants non lus. Dans le même contexte applicatif, les deux tâches sont complémentaires : la segmentation et le titrage automatique ont été utilisées dans (Joty *et al.*, 2013) pour proposer une seule partie de l'e-mail si le message aborde plusieurs thèmes. À noter que le corpus utilisé par (Joty *et al.*, 2013) possède en moyenne 2,5 thèmes par e-mail.

Notre objectif est de développer un système de titrage générique indépendant de toute information structurelle, *a priori* sur l'émission, tout en respectant la définition d'un bon titre. Nous proposons une nouvelle approche de titrage des segments de journaux télévisés obtenus automatiquement, qui exploite des informations provenant de la presse écrite. Elle consiste à apparier un segment à un article de presse du même jour traitant la même thématique, afin d'attribuer le titre de l'article à ce segment. Cette approche permet de donner à nos segments générés automatiquement des titres rédigés par des journalistes professionnels. Par conséquent, les titres remplissent les conditions nécessaires

FIGURE 7.1 – *Exemple d'un JT enrichi de la chaîne TF1. La capture d'écran a été prise à partir du site de tv-replay.fr*

pour susciter l'intérêt des lecteurs.

7.2 Positionnement du problème

7.2.1 Définition d'un titre

Le titre est annonciateur de l'information. Il permet de créer un premier lien entre le lecteur et le thème traité dans le document. Le rôle du titre est d'exprimer avec peu de mots le contenu d'un document. Pour cela, le titre doit être informatif (*i.e.* indique clairement le contenu de l'article, on dit aussi fidèle au contenu) et accrocheur (*i.e.* capte l'attention du lecteur). Bien évidement, le titre doit être aussi correct et compréhensible par le lecteur. Dans le domaine du journalisme, le titre est soigneusement mis en valeur, car c'est la partie qui donne à l'utilisateur l'envie de voir le contenu du segment ou de l'article. Cependant, il

doit remplir certains critères :

- Il doit être correct, court, clair et compréhensible. Cela passe par l'utilisation de mots simples et forts qui vont permettre de raccourcir le titre jusqu'à la forme la plus concise possible du contenu.
- Il doit donner un aperçu à la fois global et spécifique du contenu (l'essentiel en un coup d'œil). Un titre ambigu ou vague n'incite pas l'utilisateur à consulter le thème en question. D'ailleurs, dans le milieu journalistique, la devise est "écrire pour être lu et compris".

Dans (Lopez *et al.*, 2012), les titres sont regroupés en trois types :

- Type 1 : Très informatif et peu accrocheur
- Type 2 : Informatif et accrocheur
- Type 3 : Peu informatif et très accrocheur.

Le type du titre dépend essentiellement du genre du document (Lopez *et al.*, 2012). Par exemple, les titres des articles de presse doivent être de *type 2*. Ainsi, un coup d'œsuffira pour avoir une idée sur le contenu et en même temps éveiller l'intérêt et la curiosité. Les e-mails et les discussions de type *chat* nécessitent plutôt un titre de *type1*, car c'est le caractère informatif qui est le plus important. Les titres de *type 3* suscitent la curiosité du lecteur et concernent beaucoup plus les informations bizarres ou drôles.
L'ensemble de ces critères rend la tâche de titrage automatique plus complexe pour la machine.

7.2.2 Sources d'information pour la détermination d'un titre

Le contenu même de l'émission peut contenir des informations utiles à la tâche de titrage :

— *Visuel* : un changement de thème peut être accompagné par un titre vidéo. L'*OCR* (Optical Character Recognition) permet de détecter et de reconnaître les textes d'une vidéo quand ils sont incrustés.

— *Acoustique* : certains événements sonores, comme les jingles, peuvent aider à la détection de l'énoncé des titres par le présentateur.

— *Linguistique* : les mots les plus pertinents peuvent jouer le rôle d'un titre. Ainsi, une partie du discours qui contient le plus grand nombre de mots discriminants peut être aussi considérée comme un titre.

L'extraction des titres à partir du document même n'est pas toujours possible. En effet, les indicateurs visuels et acoustiques sont fortement liés aux choix éditoriaux des chaînes télévisées (le titrage incrusté n'est pas toujours disponible, le présentateur principal ne donne pas forcément en début du journal

tous les thèmes abordés, *etc.*). Le contenu linguistique de l'émission donne des titres moins informatifs, les mots clés peuvent fournir un titre ambigu et/ou demandent un effort supplémentaire de la part de l'utilisateur. De plus, l'extraction d'une portion du contenu est un problème qui dépend du système de reconnaissance automatique de la parole et du découpage en unités élémentaires.
Nous proposons dans ces travaux d'utiliser une source d'information externe, à savoir des titres d'articles de presse écrite.

7.3 Précédents travaux

Dans cette section, nous donnons un panorama général des algorithmes de titrage automatique les plus utilisés ainsi que les métriques d'évaluation auxquelles on a reconnu.

7.3.1 Méthodes de titrage automatique

Parmi les systèmes de titrage mis en place, on distingue ceux par *extraction*, ceux par *appariement* et ceux par *reformulation*.

Les méthodes dites de *titrage par extraction* consistent à extraire un ensemble de mots ou une phrase à l'aide de critères statistiques. Ces méthodes ont reçu beaucoup d'attention dans le cadre de la modélisation des thèmes [1], comme le montrent les publications récentes (Andrzejewski et Buttler, 2011), (Newman *et al.*, 2012) et (Kou *et al.*, 2015). Chaque thème latent est interprété par un ensemble de termes représentant le mieux possible son contenu. La première étape est de donner un score pour chaque mot. Par la suite, il faut choisir les mots les plus représentatifs du thème par ordre d'importance. Donc, le titrage des thèmes latents peut être vu comme étant l'indexation d'un ou plusieurs documents.

Généralement, l'estimation de ces mots nécessite un corpus de référence [2] comportant des thèmes similaires au corpus ciblé. Un nombre considérable de critères ont été proposés pour calculer le poids de chaque mot. Pour étiqueter les thèmes latents, (Andrzejewski et Buttler, 2011) ont utilisé le corpus de référence *Wikipedia* pour calculer le degré d'association entre les mots. Chaque thème est représenté par une liste de $Top10$ mots (*i.e* les dix premiers mots les plus importants du thème). Le classement des mots est généralement établi par un vote

1. La modélisation est définie dans la sous-section 1.2
2. Wikipedia est le plus utilisé

d'un ensemble de critères. Par exemple, (Zhao *et al.*, 2001) utilisent l'algorithme de *PageRank* pour effectuer un classement des mots des *tweets* lesquels selon leur importance. Dans (Lau *et al.*, 2010), les auteurs proposent une méthode pour sélectionner le mot le plus représentatif du thème à partir d'une liste de $Top10$ mots. Ceci peut être vu comme étant un reclassement des termes candidats. Pour cela, plusieurs mesures d'association entre les mots ont été exploitées.

(Hsueh et Moore, 2006) cherchent à étiqueter des réunions, pour cela un ensemble limité de labels a été défini. La tâche de titrage est considérée comme un problème de classification multi-classes. Pour chaque N-gramme candidat, quatre mesures de similarité entre les mots ont été calculées : *Log Likelihood (LL)*, *Chi-carré* (χ^2), *PMI* et le *coefficient Dice*.

L'algorithme présenté dans (Kastner et Monz, 2009) cherche la phrase la plus importante de l'article qui sera utilisée comme titre. Pour cela, les auteurs utilisent un certain nombre de critères comme le positionnement de la phrase dans le document. Cette information est considérée comme fondamentale, puisque c'est la première phrase qui donne généralement l'essentiel de l'information. Dans (Lopez *et al.*, 2011), les auteurs proposent une approche nommée *POSTIT* (TITrage par information de POSition) pour générer des titres très informatifs et peu accrocheurs. La première étape du processus consiste à extraire l'information pertinente dans les deux premières phrases du texte. Dans la deuxième étape, les syntagmes nominaux candidats au titrage sont sélectionnés. Pour cela, les auteurs utilisent l'outil *SYGFRAN* pour effectuer l'étiquetage morphosyntaxique. Un ensemble de patrons syntaxiques a été défini pour extraire les syntagmes corrects ayant une taille maximale de 9 mots. La dernière étape permet de sélectionner le syntagme nominal le plus correct selon des critères purement statistiques. Le score calculé est fondé sur la position du syntagme nominal dans le texte ainsi que sur la pertinence des termes qui le composent.

Les méthodes de titrage *par formulation* visent à produire un titre en respectant la définition d'un bon titre (informatif, accrocheur et court) ainsi que les contraintes stylistiques. Dans (Lopez *et al.*, 2012), les auteurs proposent une approche nommée *NOMIT* (Titrage par NOMInalisation) permettant de produire automatiquement des titres à la fois informatifs et accrocheurs (*i.e* de type 2). Pour que le titre soit accrocheur, il faut apporter des modifications sur la structure morphosyntaxique de l'information afin de lui attribuer une structure plus incitative en remplissant toujours les caractéristiques d'un titre pertinent. Les auteurs considèrent que les titres à base nominale sont plus incitatifs que les titres à base verbale. L'approche *NOMIT* est composée de trois étapes : l'extraction des candidats, les traitements linguistiques et la sélection du titre. La première étape permet de sélectionner les informations pertinentes à partir des deux premières phrases de l'article vérifiant le patron syntaxique suivant : forme auxiliaire conjuguée + participe passé qui sera nominalisé dans l'étape

de traitement linguistique. Par exemple, la phrase« *c'est à St-Martin que l'épicerie a été braquée* » devient *St Martin : braquage de l'épicerie*. La dernière étape permet de sélectionner le titre le plus pertinent qui s'appuie sur une validation Web (calculer la dépendance entre les termes composant les titres candidats).

Les méthodes de titrage *par appariement* cherchent des documents titrés qui traitent de la même thématique que le document cible. (Lau *et al.*, 2011) ont proposé une méthode dans laquelle les titres des articles Wikipédia sont des candidats potentiels. Le *Top*10 mots de chaque thème a été utilisé comme une requête. Ainsi, le titre du document le plus proche de la requête est considéré comme un identifiant du thème. Dans la tâche de segmentation thématique des documents audiovisuels, des notices sont parfois disponibles et la tâche de titrage consiste à associer un segment à la description d'une notice (Guinaudeau, 2011).

7.3.2 Métriques d'évaluation

Aucune métrique d'évaluation n'est spécifiquement dédiée au titrage automatique. D'ailleurs, la plupart des systèmes ont été évalué par des métriques utilisées pour le résumé automatique. Cela est essentiellement dû à la ressemblance de ces deux tâches qui visent à donner l'essentiel du contenu d'un document en peu de mots.

Dans (Jin et Hauptmann, 2000), la métrique *F-mesure* a été utilisée, en comparant les mots du titre proposés par le système à celui de la référence (l'annotation humaine), l'ordre des mots a été ignoré. En d'autres termes, la *Précision* et le *Rappel* sont calculés à partir du nombre de mots en commun et du nombre de mots constituant le titre de référence (pour le *Rappel*) ou ceux d'hypothèse (pour la *Précision*). Il faut noter que dans certains travaux comme ceux qui ont été présentés dans (Wan *et al.*, 2003), seul le *Rappel* est utilisé dans l'évaluation. (Witbrock et Mittal, 1999) et (Banko *et al.*, 2000) signalent que la métrique d'évaluation basée sur la mise en correspondance exacte entre les mots des titres de la référence et les titres automatiques présente certaines limites. Une sous-estimation est constatée (*i.e.* certains titres générés par le système sont corrects mais sont pénalisés par la métrique d'évaluation). Les auteurs expliquent ce phénomène par le fait que la mesure ne tient pas compte de l'aspect sémantique des titres. Par conséquent, l'évaluation manuelle est une nécessité.

Dans (Lopez *et al.*, 2012), les auteurs ont lancé un appel à participation sur les listes de diffusion. L'évaluateur a le choix entre quatre étiquettes : *pertinent, assez pertinent, non pertinent, ne se prononce pas*. Au niveau de l'accroche des titres, l'évaluateur a le choix entre trois étiquettes : *accrocheur, non accrocheur, ne se prononce pas*. A la fin de l'annotation, chaque titre a cinq scores : $score_{pertinent}$,

$score_{assez\_pertinent}$, $score_{non\_pertinent}$, $score_{accrocheur}$ et $score_{non\_accrocheur}$.
Les auteurs considèrent qu'un titre est pertinent si $score_{pertinent} > score_{non\_pertinent}$ et est assez pertinent si $score_{assez\_pertinent} > score_{non\_pertinent}$. Un titre est considéré accrocheur si au moins deux évaluateurs l'ont jugé accrocheur. Dans l'évaluation de l'approche *POSTIT*, les auteurs utilisent les mesures *Rappel*, *Précision* et *F-mesure* pour évaluer leurs approches.

Dans (Lau *et al.*, 2011), les auteurs utilisent l'application *Amazon Mechanical Turk* (voir la figure 7.2). Elle consiste à effectuer la tâche d'annotation par des

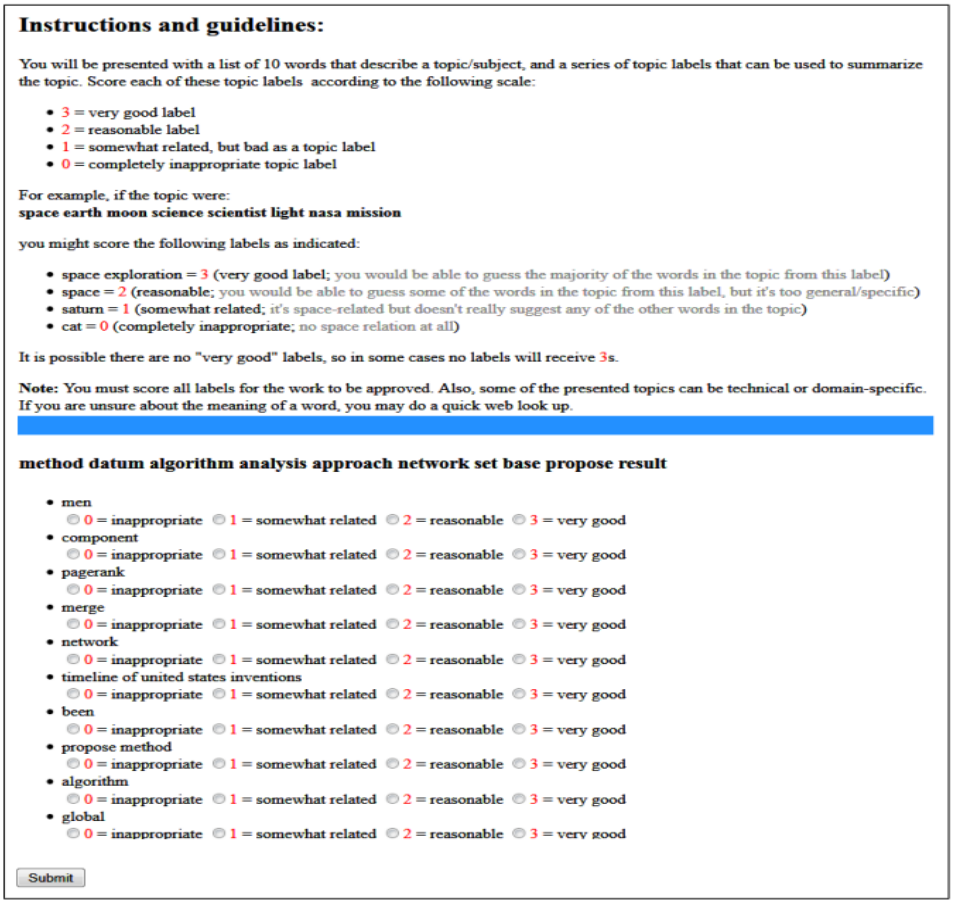


FIGURE 7.2 – *L'évaluation de titrage avec le site Amazon Mechanical Turk.*

humains contre rémunération. Un titre d'hypothèse est constitué de 10 mots. L'annotation consiste à donner, pour chaque mot, l'un de ces scores : très bon (3), raisonnable (2), lié sémantiquement au segment(1) et enfin complètement inapproprié (0). L'inconvénient est que cette méthode est coûteuse et propice à l'erreur, dans le cas où l'annotateur n'a pas de qualification pour ce genre de tâche. Concernant l'évaluation, les auteurs considèrent que la métrique Rappel/Précision n'est pas adaptée à ce genre de tâche et utilisent la métrique

nDCG (Normalized Discounted Cumulative Gain) conçue à l'origine pour la recherche d'information.

(Dorr *et al.*, 2003) utilisent la mesure *Bleu* qui a été initialement proposée pour évaluer les systèmes de traduction automatique. Cette métrique est fondée sur la comparaison des *n-grammes* de la traduction produite manuellement par des humains avec celle qui a été générée automatiquement. C'est ce principe qui a été utilisé pour comparer les titres de référence et ceux d'hypothèses.

7.4 Principe général de l'approche proposée

Notre approche de titrage intervient après la phase de segmentation thématique de JT. Elle consiste à associer à chaque segment obtenu le titre de l'article de presse du jour traitant de la même thématique. Parmi tous les articles de presse disponibles, le titre associé au segment est celui de l'article le plus proche thématiquement. Pour cela, une similarité est calculée entre le segment en question et les articles de presse du même jour. Ainsi, le titrage peut être vu comme une tâche d'ordonnancement d'une liste de titres candidats. Ce classement est effectué par le biais des valeurs de similarité. Celles-ci peuvent être aussi utiles pour effectuer un filtrage de l'ensemble des candidats. En effet, un thème abordé dans un journal télévisé n'est pas forcément traité dans la presse du web. Par conséquent, le premier document proposé prendra une valeur de similarité faible. L'application d'un seuil α permettra de régler ce genre de problème en tenant uniquement compte des documents dont on est sûr qu'ils appartiennent à la même thématique. Dans le cas d'un rejet (aucun candidat n'est retenu), on parle d'un segment *non titrable* ($\overline{T}$). Dans le cas contraire, il est considéré comme *titrable (T)*. Le titre associé au segment est celui de l'article qui maximise la similarité.
Le processus de titrage des segments thématiques à partir des articles de la presse écrite se résume en trois étapes principales, décrites dans les sous-sections suivantes.

7.4.1 Collecte des articles de presse

Cette étape consiste à récolter tous les articles de presse parus le même jour que les segments thématiques de l'émission à titrer. Nous avons choisi d'extraire les articles de presse sur la page d'accueil de *Google Actualités* qui présente de façon automatisée des informations provenant de différentes sources d'informa-

tion comme les sites Internet des journaux. [3] Ces articles traitent de domaines variés comme la politique, le sport, l'économie, *etc.* Les pages étant bruitées par des éléments non informatifs, n'extraire que le contenu utile de l'article est une nécessité dans les traitements ultérieurs. Pour notre part, nous utilisons *Boilerpipe* qui nous paraît le plus efficace. Cet outil est basé sur les arbres de décision pour déterminer la pertinence des différents blocs textuels de la page.
Nous utilisons la même base de données d'articles qui nous a servie dans l'extraction des relations sémantiques de types diachroniques (voir la sous-section 6.5.2). Pour rappel, chaque tuple (article de presse) de la base de données contient l'identifiant de l'article, le lien de la page web, le titre de l'article, la date et l'heure de la publication ainsi que l'identifiant de l'article principal.

7.4.2 Représentation vectorielle

Le calcul de la similarité entre segment et article de presse nécessite une représentation vectorielle de l'un et l'autre. Le segment thématique est issu de la transcription de la parole (contenant potentiellement des disfluences et des erreurs de reconnaissance) et l'article est composé de texte écrit dans un style journalistique. La représentation choisie doit donc être robuste à diverses sources et styles.
Nous utilisons la représentation par sac de mots qui associe à chaque document sa description vectorielle. Cette dernière contient les mots présents dans le document ainsi que leur fréquence. Cette représentation est assez simple et permet d'effectuer diverses opérations comme l'amplification des mots importants et le calcul de similarité entre les documents.
Avec l'aide du logiciel *Lia-tagg*, des pré-traitements standards (lemmatisation, filtrage des mots) ont été appliqués sur les articles de presse et les segments. Ensuite, chaque article de presse et chaque segment est représenté par une liste de mots pertinents selon la mesure *Okapi*. Nous utilisons la pondération *intra-document* pour sélectionner les mots caractérisant chaque segment thématique détecté par rapport aux autres. En revanche, nous appliquons la pondération standard pour les articles de presse. Enfin, pour une meilleure comparaison, une normalisation est réalisée relativement aux mots ayant le score le plus élevé, ce qui permet d'avoir des valeurs comprises entre 0 et 1. Finalement, un filtrage est appliqué : seuls les mots ayant un score supérieur à $+0.25$ sont conservés.

3. Le Parisien, Europe1, Francetv info, Pourquoi Docteur, Journal de l'économie, Marketing, Radio Télévision Suisse, Radio Chine Internationale, *etc.*

7.4.3 Calcul de similarité

Le but de cette étape est de calculer la similarité entre chaque couple (article, segment). Pour rappel, l'article et le segment sont représentés par la liste des mots dont le score dépasse 0.25. Le choix de la mesure appropriée à la nature de nos documents est très important. Pour cela, nous proposons de comparer les mesures *Set Jaccard, Extended Jaccard, LIN* et *Cosine* (voir le tableau 1.3). Les valeurs de similarité sont exploitées non seulement pour classer les articles mais aussi pour filtrer les résultats.

7.4.4 Métrique d'évaluation proposée

Pour une évaluation complète de notre système de structuration thématique (*i.e* l'évaluation conjointe de la segmentation thématique et le titrage), nous proposons la métrique *STER* (Segmentation and Titling Error Rate). Le processus d'évaluation se déroule en deux phases :

- Évaluation thématique de chaque segment séparément. Pour rappel, un segment est considéré comme correct si $Couv_{R \leftrightarrow H} > \gamma$ (voir la sous-section 2.4.4).
- Évaluation du titrage qui consiste à vérifier si le titre proposé par le système figure dans la liste des titres corrects.

À noter que la deuxième phase d'évaluation est applicable uniquement pour des segments corrects, pour cela nous écartons les segments erronés.

Le système de titrage mis en place peut être évalué en étudiant les différentes erreurs possibles :

- Substitution (*Sub*) : le segment est *titrable* et le titre affecté au segment n'est pas correct.
- Faux rejet (*FR*) : le segment est *titrable* et le système ne propose aucun titre.
- Fausse alarme (*FA*) : le segment est *non titrable* et le système propose un titre.

Les réponses correctes sont de deux types :

- TC : le segment S est titrable et le titre associé est correct.
- $\overline{TC}$: le segment S est non titrable et aucun titre n'est associé à ce segment.

Le *TER* (Titling Error Rate) est alors donné par :

$$TER = \frac{\#Sub + \#FR + \#FA}{\#R} \tag{7.1}$$

où $\#R$ est le nombre de segments de référence.

Un titre affecté à un segment est considéré comme bon *si et seulement si* la *segmentation thématique est jugée comme correcte* et le *titre affecté figure dans la liste des titres candidats*. Pour une évaluation complète, il suffit d'ajouter le nombre de segments d'hypothèse incorrects $H_{Err\gamma}$ comme une source d'erreur :

$$STER_\gamma = \frac{\#H_{Err\gamma} + \#Sub_\gamma + \#FR_\gamma + \#FA_\gamma}{\#R} \tag{7.2}$$

7.5 Expériences et résultats

7.5.1 Corpus et annotation

Durant la période du 10 au 16 février et avec un rythme d'une fois par heure, nous avons collecté grâce au site web *Google Actualités* une base de données de $22k$ articles. Le grand nombre d'articles de la collection et de segments de journaux télévisés rend la tâche d'annotation manuelle très longue et fastidieuse. En effet, il faut que l'annotateur évalue la potentielle association thématique entre chaque segment et les articles de la collection. Afin de réduire le nombre d'associations à évaluer, nous proposons, tout d'abord d'exploiter uniquement les articles principaux, ce qui donne en moyenne 660 articles par jour. Ainsi la collection peut être ramenée à $5.4k$ articles si l'on prend en compte uniquement les articles principaux. Ensuite, nous donnons à l'annotateur pour chaque segment, uniquement l'ensemble des articles de presse du même jour ayant au moins 2 mots en commun avec le segment considéré. Ainsi, l'annotateur a vérifié en moyenne 127 titres par segment (et non les 660 pour chaque segment).

Un titre est censé résumer le contenu du segment. Or, il peut arriver que le titre affecté ne soit pas très proche de son contenu. Par exemple, le segment qui porte sur le déploiement d'un robot sous-marin pour rechercher l'avion disparu peut avoir les hypothèses suivantes :

- « Un robot sous-marin déployé dans la zone de recherche du MH370 » : ceci correspond très bien au contenu du segment.
- « Boeing disparu de la Malaysia Airlines » : cette proposition ne couvre pas la totalité du segment.
- « Boeing meilleur qu'Airbus ? » n'a aucune relation avec l'information présentée dans le segment.

Afin d'être précis dans l'évaluation, chaque segment est associé à un ensemble de couples $E = \{\emptyset \cup (titre_1, score_1) \cup (titre_2, score_2) \cup ... \cup (titre_n, score_n)\}$. $n = 0$ signifie qu'aucun titre correspondant n'a été trouvé parmi les articles de la collection. Dans ce cas, il s'agit d'un segment non titrable, soit de type $\overline{T}$.

Afin d'évaluer le bon appariement d'un titre à un segment, nous définissons une mesure subjective qui porte essentiellement sur la corrélation entre le titre et le contenu du segment. La tâche d'annotation consiste à affecter à chaque titre un score qui indique si :

- Le titre reflète bien le contenu du segment : score 2,
- Le titre résume partiellement le segment, c'est-à-dire que le titre ne couvre pas la totalité du segment ou ne suit pas l'actualité : score 1.
- Le titre n'a pas de relation avec le segment : score 0.

Dans ce travail, uniquement les titres résumant parfaitement le contenu du segment (score 2) sont considérés comme corrects et les autres (les score 0 et 1) sont considérés comme incorrects.

Le corpus *MCS7-14* se compose de 658 segments de type T et 339 segments de type $\overline{T}$. Le tableau 7.1 représente la répartition des segments selon leur durée. Le corpus contient seulement 66, 0% de segments titrables et 34, 0% de non titrables. Cela s'explique par le fait que les journaux télévisés traitent non seulement de l'actualité du jour mais aussi des informations de société qui ne se retrouvent pas nécessairement dans les articles de presse du jour (au moins dans la même journée de diffusion de l'émission contenant le thème en question).

	Nb	Dur. Moy	T	$\overline{T}$
Seg. longs	761	131,4	467	294
Seg. courts	236	20,4	191	45
Tous les Seg.	997	105,1	658	339

***Tableau 7.1** – Répartition des segments titrables T et non titrables $\overline{T}$ par rapport à la durée des segments (courts et longs) du corpus MCS7-14.*

7.5.2 Évaluation sur les segments de référence

Dans un premier temps, nous évaluons le titrage sur les segments de référence (définis manuellement) avec différentes mesures de similarité. La figure 7.3 illustre la performance du système sous forme d'une courbe *ROC* (substitution et faux rejet en fonction de fausse alarme). Les résultats ont été obtenus en faisant varier le seuil α appliqué sur les valeurs de similarité. La première observation indique que les mesures pondérées (*i.e. Lin, Exteded_Jaccard* et *Cosine*) donnent de meilleurs résultats que le *Set_Jaccard*. En effet, la pondération donne plus de poids aux mots importants et pénalise les mots les moins représentatifs. Ceci a une influence directe sur le calcul de la cohésion et donc sur le classement des titres candidats. La mesure *Cosine* donne les meilleures performances par

rapport à *Extended_Jaccard* et *LIN*. Par la suite, les performances sont données pour la mesure *Cosine*.

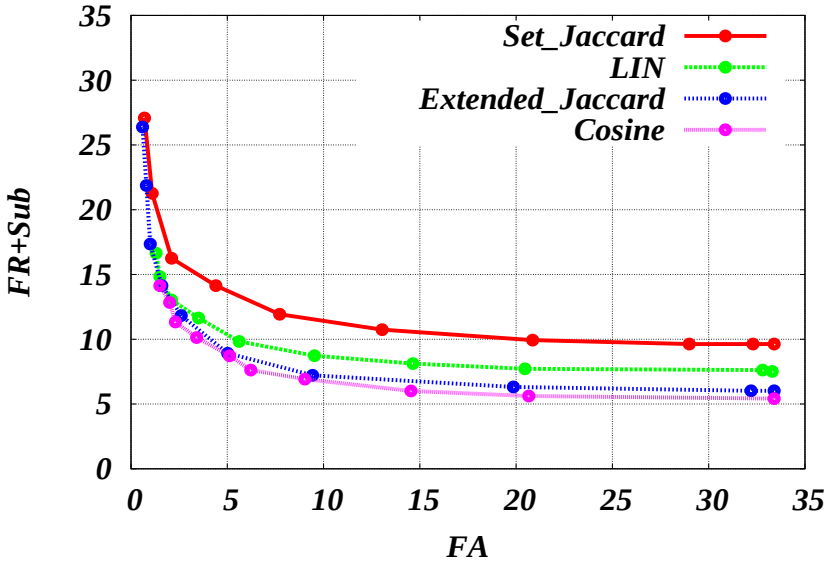


FIGURE 7.3 – *TER pour les différentes mesures de similarité*

Le tableau 7.2 donne la qualité de titrage sur les segments courts et longs. Dans sa globalité, le système conduit à de très bonnes performances. Entre les erreurs de substitution, les fausses alarmes et le rejet correct, la qualité du titrage est de l'ordre de 88.2%. Notre système a la capacité de bien titrer non seulement les segments longs (représentés par un nombre important de termes) mais aussi les segments courts. Et cela même si on observe de mauvaises performances en terme de nombre de segments correctement titrés pour les segments courts par rapport aux segments longs. La plupart des segments courts correctement titrés correspondent à des thèmes qui ont été traités en force dans les articles de presse. Cependant, dans le journal d'information, ils sont traités assez rapidement .

Cosine	*TER*	*Sub*	*FA*	*FR*
Seg. longs	11.3	3.4	3.4	4.5
Seg. courts	13.6	5,5	5.1	3,0
Tous les seg.	11.8	3,9	3.8	4,1

Tableau 7.2 – *Performance du système de titrage sur les segments de référence du corpus MCS7-14. La similarité est calculée avec la mesure Cosine.*

Une analyse des erreurs a mis en évidence deux catégories principales d'erreurs : (1) celles qui proviennent de fils d'actualités dont le contenu évolue à court terme, (2) celles qui proviennent d'informations traitant de thématiques

proches mais légèrement différentes. Le tableau 7.3 fournit quelques exemples d'erreurs.

	Titre correct	Titre erroné	Commentaires
1	Sotchi : la France attend sa première médaille	JO : Martin Fourcade, premier médaille d'or français.	Dépend de l'heure à laquelle l'information est donnée. Ce genre d'informations peut changer plusieurs fois dans la journée.
2	Le perchiste Renaud Lavillenie accueilli en héros à Roissy.	Saut à la perche : Lavillenie incertain pour les mondiaux en salle.	
3	Crues : trois départements en vigilance orange.	Tempête : Quelque 10.000 foyers toujours privés d'électricité en Bretagne	Des informations traitent des thématiques proches
4	Internet : une campagne pour dire non au harcèlement des ados.	Grâce à Facebook, un père va pouvoir se remémorer le passé de son fils décédé.	

Tableau 7.3 – *Exemples d'erreurs de titrage sur des segments de référence du corpus MCS7-14.*

7.5.3 Évaluation sur les segments automatiques

Nous utilisons le système de segmentation dans sa configuration optimale (*i.e.* la combinaison linéaire entre *Word2vec* et *Vitanyi*). Avec cette configuration, la *F-mesure* est égale à 76.3% avec un *Rappel* de 73.6% et une *Précision* de 79.1% (pour plus de détails voir la section 6.6.2). L'évaluation au niveau des segments est donnée dans le tableau 7.4. Les résultats ont été obtenus en faisant varier le seuil de couverture γ.
Nous remarquons une différence de performance entre les segments longs et les segments courts. En effet, contrairement aux segments courts, le système est nettement meilleur sur les segments longs. Ces derniers sont plus importants à extraire que les segments courts (qui ne durent que quelques secondes et correspondent généralement à des brèves).

SER_γ (%)	$\gamma = 80\%$	$\gamma = 85\%$	$\gamma = 90\%$
Seg. longs	21.0	23.9	30.4
Seg. courts	67.8	72.5	78.4
Tous les seg. / nombre	30.3	33.7	41.7
Tous les seg. / durée	21.9	24.9	30.6

Tableau 7.4 – *Performance du système en terme de nombre de segments détectés.*

Le tableau 7.5 donne le comportement de notre système de titrage sur des segments automatiques. Avec une couverture minimale, le nombre de segments considérés comme corrects est important ce qui donne des valeurs faibles de $H_{Err\gamma}$ et par conséquent impacte la métrique $STER_\gamma$. De l'autre côté, avec une couverture maximale le système propose moins de segments corrects et donc une valeur assez élevée de $H_{Err\gamma}$ et $STER_\gamma$.
Pour le titrage, le *TER* (voir l'équation 7.1) est respectivement de 8.6, 7.8 et 7.0 pour une couverture de 80, 85 et 90. Cela signifie que plus la qualité de segmentation est bonne, plus les erreurs de titrage diminuent ce qui permet au système

γ	$STER_\gamma$	$H_{Err\gamma}$	Sub_γ	FA_γ	FR_γ
80	38.9	30.3	2.6	4.5	1.5
85	41.5	33.7	2.5	3.9	1.4
90	47.2	40.2	2.3	3.5	1.2

Tableau 7.5 – *Taux d'erreur prenant en compte à la fois la segmentation et le titrage.*

de garder la capacité d'affecter de bons titres aux segments. En effet, une faible valeur de couverture correspond à l'un des deux phénomènes suivants :

- Le segment d'hypothèse couvre partiellement le segment de réference (voir le segment H_4 de la figure 2.11). Par conséquent, la liste des mots décrivant le contenu du segment est incomplète.
- Le segment d'hypothèse couvre partiellement et/ou totalement deux segments de réference (voir le segment H_1 de la figure2.11). La liste des mots décrivant le contenu du segment est alors bruitée.

Pour comparer la robustesse du titrage sur des segments générés manuellement et automatiquement, nous évaluons le titrage sur le même ensemble. Pour cela, nous écartons les segments mal déterminés par le système $(Couv_{R\leftrightarrow H} < 85)$. Avec les 644 segments corrects, le *TER* est respectivement de 10.6% ($Sub_{85} = 3.3$, $FA_{85} = 1.7$, $FR_{85} = 5.6$) et 12.1% ($Sub_{85} = 3.9$, $FA_{85} = 2.2$, $FR_{85} = 6.0$) pour les segments manuels et automatiques. Ces résultats mettent en évidence que la bonne détermination des segments aidera le système à choisir les bons titres. Cependant, les performances restent acceptables avec les segments automatiques, ce qui illustre bien la robustesse de notre système de titrage.

Enfin, nous avons appliqué notre algorithme de titrage sur le corpus *MCS5-15*, les performances sont données dans le tableau 7.6.
En analysant les résultats, il s'avère qu'un nombre important d'erreurs de type substitution ont été produites pour les deux segmentations (manuelle et automatique). Cela revient essentiellement aux particularités du corpus. En effet, durant la période du 26/01/2015 et 27/01/2015, les informations ont été beaucoup plus focalisées sur le chapitre des élections Grecques qui a été coupé en

γ	Nb. Seg	$H_{Err\gamma}$	Sub_γ	FA_γ	FR_γ
Seg. manuelle	297	0.0	10.8	4.6	3.2
Seg. automatique	187	37.0	9.1	3.1	1.8

***Tableau 7.6** – Évaluation du titrage sur les segments de référence et automatiques du corpus MCS5-15.*

une succession de thèmes dont chacun porte sur une information bien particulier (législatives Grèce, réaction de la zone euro, dettes Grecque, *etc*). Le risque de confusion dans ce genre d'informations est beaucoup plus important. Les mêmes erreurs constatées dans le corpus *MCS7-14* ont été reproduites sur les émissions de *MCS5-15*, quelques exemples sont donnés dans le tableau 7.7.

	Titre correct	Titre erroné
1	Hervé Gourdel. Son corps est arrivé en France.	Le corps d'Hervé Gourdel rapatrié en France ce lundi.
2	L'article traite la hausse des cyclistes en ville (pas de titre dans la base).	Morts sur les routes : 26 mesures pour enrayer la hausse.

***Tableau 7.7** – Exemples d'erreurs de titrage sur des segments de référence du corpus MCS5-15.*

7.6 Conclusion

Dans ce chapitre, nous avons décrit notre système de titrage automatique des segments. Après l'étape de segmentation thématique appliquée à des sorties d'un système de reconnaissance de la parole, les segments obtenus sont titrés. Le titrage consiste à apparier le segment à un article de presse traitant du même thème. Le titre associé au segment est celui de l'article qui maximise la similarité entre le segment et les articles du jour. Une métrique d'évaluation mesurant conjointement la qualité de la segmentation et du titrage a été proposée. Les résultats obtenus montrent que les erreurs de segmentation restent prépondérantes dans le processus. Le titrage donne de bons résultats et est robuste aux petites imprécisions de la segmentation. Comme perspective à ce travail, il est envisagé d'étudier l'interaction entre ces deux tâches afin d'améliorer la qualité du système de segmentation à partir du titrage, notamment pour les segments courts.

Conclusion générale et perspectives

Conclusion

Le thème auquel nous nous intéressons est celui de la structuration thématique des journaux télévisés. Le système que nous avons développé comporte deux tâches complémentaires : la *segmentation thématique* et *le titrage*. La segmentation thématique permet de découper l'émission en passages thématiquement homogènes. Ces derniers sont identifiés par des labels anonymes. Le titrage automatique permet de donner un en-tête pour chaque segment détecté. Ce duo donne la possibilité aux utilisateurs de naviguer plus rapidement à l'aide d'une table de thèmes contenant le titre, l'instant de début, et l'instant de fin de chaque thème. L'importance de la structuration thématique ne se restreint pas à des fins de navigation. Elle est considérée comme une première étape vers d'autres applications comme la recherche d'information et le résumé automatique.

Trois catégories d'indices peuvent être exploitées pour segmenter des journaux d'informations télévisés : indices lexicaux, acoustiques et visuels (segmentation thématique multimodale). Bien évidemment, la combinaison de ces indices est en règle générale profitable, mais le système peut être fortement lié aux règles éditoriales de chaque chaîne. Étant donné que notre objectif était de développer un système de segmentation thématique applicable sur n'importe quel journal d'information, nous avons privilégié les indices lexicaux et les informations liées aux locuteurs.

Dans ces travaux, nous avons mis l'accent sur l'enrichissement de la représentation vectorielle de l'émission. Nous avons d'abord proposé deux approches de pondérations intra-document : *itérative* et *à base d'informations structurelles*. Les deux propositions essaient de découper l'émission en N chunks de différentes tailles. Le principe général du calcul de pondération utilisé en recherche d'information est appliqué pour l'estimation des poids des mots prononcés dans le journal, en considérant que les chunks sont les documents.

Nous avons ensuite introduit la notion de la *cohésion de la parole*. La similarité entre les différentes parties de l'émission est calculée à partir de la distribution

des mots et des locuteurs. L'idée de base est de considérer l'identifiant du locuteur comme un terme au même titre que chaque mot, c'est-à-dire que chaque groupe de souffle est représenté non seulement par les mots prononcés mais également par l'indice du locuteur correspondant. Une frontière potentielle est valide si la distribution conjointe des mots et des locuteurs diffère suffisamment de part et d'autre de la frontière.

Le calcul de la cohésion a été également renforcé par les relations sémantiques entre les mots. Ces dernières sont estimées à partir d'articles de presse du même jour que l'émission ciblée (données *DiaNews*). L'idée est de mettre en évidence les liens thématiques entre les mots de l'actualité comme (*Sotchi* et *JO*), (*Sotchi* et *médaille*), *etc.* Les relations ont été estimées par l'outil *Word2vec* basé sur la représentation vectorielle des mots et la mesure *NWD* qui repose elle sur la Théorie de l'Information.

Afin d'évaluer nos travaux, nous nous sommes intéressés aux métriques. Celle de l'état de l'art évaluent la performance du système en comparant le positionnement des frontières thématiques de référence avec celles proposées automatiquement. Nous proposons deux nouvelles métriques d'évaluation pour la segmentation thématique nommées *CouvN* et *CouvD*. La mesure *CouvN* est basée sur le nombre de segments corrects, tandis que *CouvD* repose sur la durée des segments corrects. L'avantage d'une évaluation au niveau du segment réside dans le fait que les scores calculés sont facilement interprétables. De plus, ils reflètent la performance du système au niveau de chaque segment séparément ainsi qu'au niveau de l'émission globale.

En ce qui concerne le titrage automatique, nous avons proposé une approche par appariement. En pratique, le système cherche pour chaque segment de l'émission les articles de presse traitant du même thème. Le titre associé au segment est celui de l'article le plus proche thématiquement. Pour cela, une similarité est calculée entre le segment en question et les articles de presse du même jour. Une nouvelle métrique *STER* a été proposée pour évaluer conjointement la segmentation thématique et le titrage automatique.

Les améliorations que nous avons apportées ne sont pas limitées à la tâche de segmentation thématique. Elles peuvent avoir une portée beaucoup plus large dans de nombreux contextes d'utilisation. Par exemple, la pondération *intra-document* peut être utilisée dans le résumé automatique.

Perspectives

Un certain nombre de perspectives peuvent être envisagées afin de poursuivre les travaux présentés dans cette thèse. Nous avons notamment dégagé

trois pistes d'amélioration :

1. Valider nos contributions sur d'autres algorithmes de segmentation thématique. Cette perspective est facile à mettre en œuvre, car nos propositions mettent l'accent sur la représentation vectorielle. Cette dernière est une étape indispensable pour n'importe quel algorithme de segmentation thématique basé sur le calcul de la cohésion lexicale.
2. L'intégration d'étiquettes identifiant les locuteurs à la distribution de mots apporte un gain de qualité à la segmentation. Cela nous amène à considérer que l'ajout d'étiquettes provenant d'autres informations comme une option potentiellement bénéfique à notre travail. Afin d'espérer une amélioration de la segmentation, une étiquette doit alors respecter deux propriétés :
   — Discriminance : deux groupes de souffle appartenant à des thèmes différents ne partagent pas la même étiquette.
   — Répétition : une même étiquette correspond à plusieurs groupes de souffle.

   Nous allons à présent nous attarder davantage sur la seconde piste en l'illustrant par des exemples.
   Exemple 1 étiquette « classe thématique »
   L'actualité est traitée dans plusieurs canaux comme la presse écrite, les journaux d'information, twitter, *etc.* Nous proposons d'utiliser *Google Actualités* qui permet de récolter et de regrouper au sein d'un même cluster les articles d'information qui traitent d'un même thème et renvoient des contenus web. Chaque *cluster* est représenté par un seul article. L'exemple illustré dans la figure 6.3 contient trois classes thématiques. L'idée est d'ajouter pour chaque groupe de souffle l'indice du *cluster* le plus proche thématiquement. Pour cela, la similarité est calculée pour chaque groupe de souffle avec le représentant de chaque *cluster*.
   Exemple 2 étiquette « domaine thématique »
   Un journal d'information aborde des domaines variés comme la politique, le sport, le cinéma, la santé, *etc.* Affecter pour chaque groupe de souffle son domaine peut enrichir beaucoup plus la représentation vectorielle de l'émission. Dans l'exemple du tableau 7.8, la représentation vectorielle des groupes de souffle est composé de la distribution de locuteurs, de la classe thématique, de la distribution des mots et du domaine thématique.
3. La plupart des systèmes de segmentation thématique traitent les émissions séparément, c'est-à-dire indépendamment les unes des autres. Or, au cours d'une même journée, certains thèmes peuvent être traités dans plusieurs émissions. Les informations apportées par la segmentation d'une émission pourraient ainsi guider la segmentation d'une autre émission de

GS_1	locuteur1	C1	$mot_3\ mot_4\ mot_{18}\ mot_{71}$	politique
GS_2	locuteur1	C1	$mot_9\ mot_4\ mot_{90}\ mot_{18} mot_{14}$	politique
	...	...	...	...
GS_{m-1}	locteur13	C2	$mot_{17}\ mot_{66}\ mot_{355}\ mot_{45}\ mot_{39}\ mot_4$	sport
GS_m	locteur13	C2	$mot_{355}\ mot_{69}\ mot_{66}\ mot_{45}$	sport

***Tableau 7.8** – Représentation vectorielle des groupes de souffle (GS) contenant plusieurs informations*

la même journée. Cette dernière segmentation peut également apporter des précisions en vue d'améliorer la première segmentation. Ceci nous amène à penser qu'une *segmentation croisée (ou cross-show segmentation)*, qui au lieu de traiter les émissions indépendamment, considérerait des groupes d'émissions *a priori* semblables (par exemple les émissions d'une même journée) serait la bienvenue. Ainsi, la segmentation croisée, résultant de l'adaptation d'un système de segmentation classique appliqué à un groupe d'émissions, permettrait à la fois de guider la segmentation courante mais aussi de corriger les incohérences relevées dans les autres. Le principe de la segmentation croisée est illustré dans la figure 7.4.

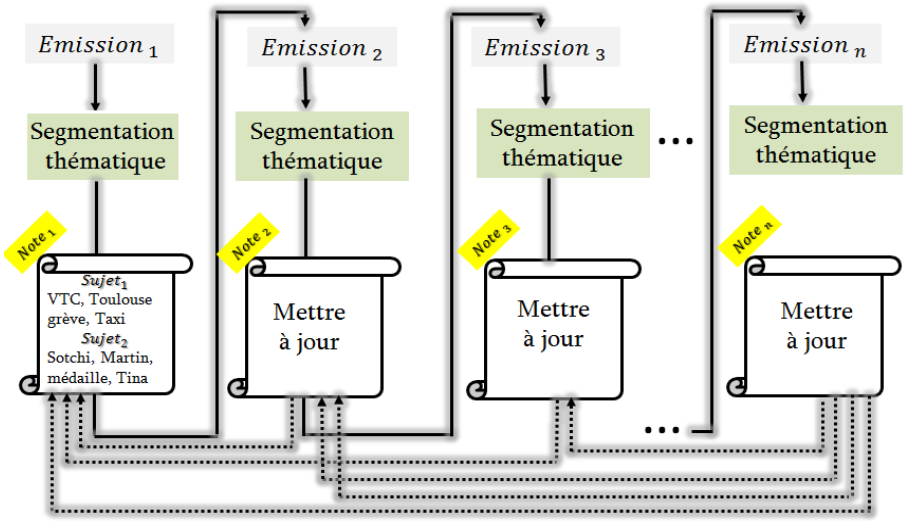


FIGURE 7.4 – *Le principe de la segmentation croisée.*

Il nous semble que les perspectives décrites ici sont peu contraignantes à mettre en œuvre et augmenteraient encore la fiabilité notre système de segmentation thématique.

Liste de publications

Abdessalam Bouchekif, Géraldine Damnati, Delphine Charlet, Nathalie Camelin and Yannick Estève. **Title assignment for automatic topic segments in TV Broadcast News**. *In 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai (China).

Abdessalam Bouchekif, Géraldine Damnati, Yannick Estève, Delphine Charlet and Nathalie Camelin. **Diachronic Semantic Cohesion for Topic Segmentation of TV Broadcast News**. *In 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden (Germany).

Delphine Charlet, Géraldine Damnati, Abdessalam Bouchekif and Ameur Douib. **Fusion of Speaker and Lexical Information for Topic Segmentation : A co-segmentation Approach**. *In 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane (Australia).

Abdessalam Bouchekif, Géraldine Damnati, Nathalie Camelin, Yannick Estève and Delphine Charlet. **Segmentation et titrage automatique de journaux télévisés**. *In 22e Conférence du Traitement Automatique des Langues Naturelles (TALN 2015)*, Caen (France).

Abdessalam Bouchekif, Géraldine Damnati and Delphine Charlet. **Speech Cohesion for Topic Segmentation of Spoken Contents**. *In 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, Singapore, 14-18 September 2014.

Abdessalam Bouchekif, Géraldine Damnati and Delphine Charlet. **Intra-content Term Weighting for Topic Segmentation**. *In 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence (Italy).

Abdessalam Bouchekif, Géraldine Damnati and Delphine Charlet. **Exploitation de la distribution des locuteurs pour la segmentation thématique de journaux télévisés**. *In 30^e Journées d'Etudes sur la Parole (JEP 2014)*, Le Mans (France).

Abdessalam Bouchekif, Géraldine Damnati and Delphine Charlet. **Complementarity of Lexical Cohesion and Speaker Role Information for Story Segmentation of French TV Broadcast News**. *International Conference on Statistical Language and Speech Processing (SLSP 2013). Lecture Notes in Computer Science 7978, Springer 2013, ISBN 978-3-642-39592-5*, Tarragona (Spain).

Abdessalam Bouchekif, Géraldine Damnati and Delphine Charlet. **Segmentation thématique : processus itératif de pondération intra-contenu**. *In* 20^{e} *Conférence du Traitement Automatique des Langues Naturelles (TALN 2013)*, Sables d'Olonne (France).

Liste des illustrations

Liste des tableaux

Bibliographie

ALLAUZEN, A. et GAUVAIN, J.-L. (2005). Diachronic vocabulary adaptation for broadcast news transcription. *In INTERSPEECH*, pages 1305–1308.

AMINI, M.-R. et GAUSSIER, É. (2013). *Recherche d'Information - applications, modèles et algorithmes*. Eyrolles.

ANDRZEJEWSKI, D. et BUTTLER, D. (2011). Latent topic feedback for information retrieval. *In KDD*, pages 600–608.

BANKO, M., MITTAL, V. O. et WITBROCK, M. J. (2000). Headline generation based on statistical translation. *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325.

BARRAS, C., ZHU, X., MEIGNIER, S. et GAUVAIN, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1505–1512.

BÉCHET, F., BENDRIS, M., CHARLET, D., DAMNATI, G., FAVRE, B., ROUVIER, M., AUGUSTE, R., BIGOT, B., DUFOUR, R., FREDOUILLE, C., LINARÈS, G., MARTINET, J., SENAY, G. et TIRILLY, P. (2015). Identification de personnes dans des flux multimédia. *In CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015.*, pages 239–251.

BEEFERMAN, D., BERGER, A. et LAFFERTY, J. (1997). Text segmentation using exponential models. *In In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 35–46.

BEEFERMAN, D., BERGER, A. L. et LAFFERTY, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2013a). Complementarity of lexical cohesion and speaker role information for story segmentation of french TV broadcast news. *In Statistical Language and Speech Processing, Tarragona, Spain*, pages 51–61.

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2013b). Segmentation thématique : processus itératif de pondération intra-contenu. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 739–746, Les Sables d'Olonne, France.

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2014a). Intra-content term weighting for topic segmentation. *In 39th IEEE International Conference on Acoustics, Speech and Signal Processing*.

BOUCHEKIF, A., DAMNATI, G. et CHARLET, D. (2014b). Speech cohesion for topic segmentation of spoken contents. *In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 1890–1894.

BOUCHEKIF, A., DAMNATI, G., CHARLET, D., CAMELIN, N. et ESTÈVE, Y. (2016). Title assignment for automatic topic segments in tv broadcast news. *In 41th IEEE International Conference on Acoustics, Speech and Signal Processing ,Shanghai, China, March 20-25 , 2016*, pages 22–29.

BOUCHEKIF, A., DAMNATI, G., ESTÈVE, Y., CHARLET, D. et CAMELIN, N. (2015). Diachronic semantic cohesion for topic segmentation of TV broadcast news. *In INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2932–2936.

CHAISORN, L., CHUA, T. et LEE, C. (2003). A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208.

CHARLET, D., BARRAS, C. et LIÉNARD, J.-S. (2013). Impact of overlapping speech detection on speaker diarization for broadcast news and debates. *In ICASSP*, pages 7707–7711.

CHARLET, D., DAMNATI, G., BOUCHEKIF, A. et DOUIB, A. (2015a). Fusion of speaker and lexical information for topic segmentation : A co-segmentation approach. *In 40th IEEE International Conference on Acoustics, Speech and Signal Processing*.

CHARLET, D., DAMNATI, G. et TRIONE, J. (2015b). News talk-show chaptering with journalistic genres. *In 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1368–1372.

CHOI, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33.

CHOI, F. Y. Y., WIEMER-HASTINGS, P. et MOORE, J. (2001). Latent semantic analysis for text segmentation. *In In Proceedings of EMNLP*, pages 109–117.

CHURCH, K. W. et HANKS, P. (1989). Word association norms, mutual information and lexicography. *In 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83.

CLAVEAU, V. et LEFÈVRE, S. (2011). Topic segmentation of tv-streams by mathematical morphology and vectorization. *In INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 1105–1108.

CURRAN, J. R. et MOENS, M. (2002). Improvements in automatic thesaurus extraction. *In Proceedings of the workshop on Workshop On Unsupervised Lexical Acquisition, Philadelphia*, pages 59–66.

DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W. et HARSHMAN, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

DIAS, G., ALVES, E. et LOPES, J. G. P. (2007). Topic segmentation algorithms for text summarization and passage retrieval : An exhaustive evaluation. *In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 1334–1339.

DORR, B., ZAJIC, D. et SCHWARTZ, R. (2003). Hedge trimmer : A parse-and-trim approach to headline generation. *In Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5*, HLT-NAACL-DUC '03.

DUMONT, E. et QUÉNOT, G. (2012). Automatic story segmentation for TV news video using multiple modalities. *Int. J. Digital Multimedia Broadcasting*, 2012: 732514 :1–732514 :11.

DUPUY, G., MEIGNIER, S., DELÉGLISE, P. et ESTÈVE, Y. a. (2014). Recent improvements on ilp-based clustering for broadcast news speaker diarization. *In Odyssey 2014 : The Speaker and Language Recognition Workshop*, Joensuu (Finland).

EISENSTEIN, J. et BARZILAY, R. (2008). Bayesian unsupervised topic segmentation. *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

FAVRE, B., DAMNATI, G., BÉCHET, F., BENDRIS, M., CHARLET, D., AUGUSTE, R., AYACHE, S., BIGOT, B., DELTEIL, A., DUFOUR, R., FREDOUILLE, C., LINARÈS, G., MARTINET, J., SENAY, G. et TIRILLY, P. (2013). PERCOLI : A person identification system for the 2013 REPERE challenge. *In Proceedings of the First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22-23, 2013*, pages 55–60.

FERRET, O. (2002). Using collocations for topic segmentation and link detection. *In 19th International Conference on Computational Linguistics, COLING , Taipei, Taiwan*.

FOURNIER, C. (2013). Evaluating text segmentation using boundary edit distance. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712.

FREDOUILLE, C. et CHARLET, D. (2014). Analysis of i-vector framework for speaker identification in tv-shows. *In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 71–75. ISCA.

FUSAYASU, Y., TANAKA, K., TAKIGUCHI, T. et ARIKI, Y. (2015). Word-error correction of continuous speech recognition based on normalized relevance distance. *In Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1257–1262. AAAI Press.

GALLEY, M., MCKEOWN, K., FOSLER-LUSSIER, E. et JING, H. (2003). Discourse segmentation of multi-party conversation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 562–569.

GAUSSIER, E. et YVON, F., éditeurs (2011). *Modèles statistiques pour l'accès à l'information textuelle*. Hermès, Paris.

GAUVAIN, J.-L., LAMEL, L. et ADDA, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, 37(1):89–108.

GEORGESCUL, M., CLARK, A. et ARMSTRONG, S. (2006a). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 144–151.

GEORGESCUL, M., CLARK, A. et ARMSTRONG, S. (2006b). Word distributions for thematic segmentation in a support vector machine approach. Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), pages 101–108.

GUINAUDEAU, C. (2011). *Structuration automatique de flux télévisuels*. Thèse de doctorat, Institut National des Sciences Appliquées de Rennes, France.

GUINAUDEAU, C., GRAVIER, G. et SÉBILLOT, P. (2010). Improving asr-based topic segmentation of TV programs with confidence measures and semantic relations. *In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1365–1368.

GUINAUDEAU, C., GRAVIER, G. et SÉBILLOT, P. (2012). Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech & Language*, 26(2):90–104.

GUINAUDEAU, C. et HIRSCHBERG, J. (2011). Accounting for prosodic information to improve asr-based topic tracking for TV broadcast news. *In INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 1401–1404.

HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64.

HSU, W. H., KENNEDY, L. S., fu CHANG, S., FRANZ, M. et SMITH, J. R. (2005). Columbia-ibm news video story segmentation in trecvid 2004. Rapport technique, The 4th International Conference on Image and Video Retrieval (CIVR.

HSUEH, P. et MOORE, J. D. (2006). Automatic topic segmentation and labeling in multiparty dialogue. pages 98–101.

HUET, S. (2007). *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole*. Thèse de doctorat, Université Rennes 1.

ILLINA, I., FOHR, D. et LINARES, G. (2014). Proper name retrieval from diachronic documents for automatic speech transcription using lexical and temporal context. *In Workshop on Speech, Language and Audio in Multimedia*.

JACCARD, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 547-579.

JANIN, A., BARON, D., EDWARDS, J., ELLIS, D., GELBART, D., MORGAN, N., PESKIN, B., PFAU, T., SHRIBERG, E., STOLCKE, A. et WOOTERS, C. (2003). The ICSI meeting corpus. *In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 364–367.

JIN, R. et HAUPTMANN, A. G. (2000). Title generation for spoken broadcast news using a training corpus. *In Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, pages 680–683.

JOTY, S., CARENINI, G. et NG, R. (2013). Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.

JOUSSE, V. (2011). *Named identification of speakers : using audio signal and rich transcription*. Thèse de doctorat, Université du Maine.

KASTNER, I. et MONZ, C. (2009). Automatic single-document key fact extraction from newswire articles. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 415–423.

KAZANTSEVA, A. et SZPAKOWICZ, S. (2012). Topical segmentation : a study of human performance and a new measure of quality. *In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 211–220.

KERN, R. et GRANITZER, M. (2009). Efficient linear text segmentation based on information retrieval techniques. *In Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '09, pages 167–171.

KOU, W., LI, F. et BALDWIN, T. (2015). Automatic labelling of topic models using word vectors and letter trigram vectors. *In Information Retrieval Technology - 11th Asia Information Retrieval Societies Conference, Australia*, Lecture Notes in Computer Science, pages 253–264.

KÅGEBÄCK, M., MOGREN, O., TAHMASEBI, N. et DUBHASHI, D. (2014). Extractive summarization using continuous vector space models. Gothenburg, Sweden.

LABADIÉ, A. (2008). *Segmentation thématique de texte linéaire et non-supervisée*. Theses, Université Montpellier II - Sciences et Techniques du Languedoc.

LAMPRIER, S., AMGHAR, T., LEVRAT, B. et SAUBION, F. (2007). On evaluation methodologies for text segmentation algorithms. *In 19th IEEE International Conference on Tools with Artificial Intelligence Patras, Greece*, pages 19–26.

LAN, G. L., MEIGNIER, S., CHARLET, D. et DELÉGLISE, P. (2016). Speaker diarization with unsupervised training framework. *In IEEE International Conference on Acoustics, Speech and Signal Processing*.

LAU, J. H., GRIESER, K., NEWMAN, D. et BALDWIN, T. (2011). Automatic labelling of topic models. *In ACL*, pages 1536–1545.

LAU, J. H., NEWMAN, D., KARIMI, S. et BALDWIN, T. (2010). Best topic word selection for topic labelling. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 605–613.

LECORVÉ, G., GRAVIER, G. et SÉBILLOT, P. (2008). An unsupervised web-based topic language model adaptation method. *In IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.*, pages 5081–5084.

LIN, D. (1998). An information-theoretic definition of similarity. *In* SHAVLIK, J. W., éditeur : *Proceedings of the Fifteenth International Conference on Machine*

Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998, pages 296–304. Morgan Kaufmann.

LOPEZ, C., PRINCE, V. et ROCHE, M. (2011). Automatic titling of articles using position and statistical information. *In Recent Advances in Natural Language Processing, Hissar, Bulgaria*, pages 727–732.

LOPEZ, C., PRINCE, V. et ROCHE, M. (2012). NOMIT : automatic titling by nominalizing. *In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June Montréal, Canada*, pages 274–283. The Association for Computational Linguistics.

LUHN, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

MALIOUTOV, I. et BARZILAY, R. (2006). Minimum cut model for spoken lecture segmentation. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Sydney, Australia.

MIKOLOV, T., CHEN, K., CORRADO, G. et DEAN, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. et DEAN, J. (2013b). Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

MISRA, H., YVON, F., JOSE, J. et CAPPE, O. (2009). Text segmentation via topic modeling : an analytical study. *In 18th ACM CIKM*.

MORIN, F. et BENGIO, Y. (2005). Hierarchical probabilistic neural network language model. *In* COWELL, R. G. et GHAHRAMANI, Z., éditeurs : *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252.

NASR, A., BÉCHET, F., REY, J., FAVRE, B. et ROUX, J. L. (2011). MACAON an NLP tool suite for processing word lattices. *In The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, USA - System Demonstrations*, pages 86–91. The Association for Computer Linguistics.

NEWMAN, D., KOILADA, N., LAU, J. H. et BALDWIN, T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. *In COLING*, pages 2077–2092.

NIE, X., FENG, W., WAN, L. et XIE, L. (2013a). Measuring semantic similarity by contextualword connections in chinese news story segmentation. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8312–8316.

NIE, X., FENG, W., WAN, L. et XIE, L. (2013b). Measuring semantic similarity by contextualword connections in chinese news story segmentation. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8312–8316.

PASSONNEAU, R. J. et LITMAN, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

PENNINGTON, J., SOCHER, R. et MANNING, C. D. (2014). Glove : Global vectors for word representation. *In International Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

PEVZNER, L. et HEARST, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36.

PRINCE, V. et LABADIÉ, A. (2007). Text segmentation based on document understanding for information retrieval. *In Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27-29, 2007, Proceedings*, pages 295–304.

REYNAR, J. C. (1994). An automatic method of finding topic boundaries. *In 32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings.*, pages 331–333.

RICHMOND, K., SMITH, A. et AMITAY, E. (1997). Detecting subject boundaries within text : A language independent statistical approach. *In Second Conference on Empirical Methods in Natural Language Processing*, pages 47–54.

ROBERTSON, S. E. et JONES, K. S. (1976). Relevance weighting of search terms. *JASIS*, 27(3):129–146.

ROSE, L. S. et CHANDRAN, K. (2012). Normalized web distance based web query classification. *Journal of Computer Science*, 8(5):804.

ROSENBERG, A. et HIRSCHBERG, J. (2006). Story segmentation of broadcast news in english, mandarin and arabic. *In Proceedings of the Human Language Technology Conference of the NAAC*, NAACL-Short '06, pages 125–128.

ROSENBERG, A., SHARIFI, M. et HIRSCHBERG, J. (2007). Varying input segmentation for story boundary detection in english, arabic and mandarin broadcast news. *In INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2589–2592.

SAKAHARA, M., OKADA, S. et NITTA, K. (2014). Domain-independent unsupervised text segmentation for data management. *In 2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, pages 481–487.

SALTON, G., WONG, A. et YANG, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

SCAIANO, M. et INKPEN, D. (2012). Getting more from segmentation evaluation. *In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 362–366.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

SHRIBERG, E., DHILLON, R., BHAGAT, S., ANG, J. et CARVEY, H. (2004). The icsi meeting recorder dialog act *mrda* corpus. Rapport technique.

SHRIBERG, E., STOLCKE, A., HAKKANI-TÜR, D. Z. et TÜR, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

SIENCNIK, S. K. (2015). Adapting word2vec to named entity recognition. *In Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 239–243.

SILBER, G. et MCCOY, K. (2000). An efficient text summarizer using lexical chains. *In INLG 2000 - Proceedings of the First International Natural Language Generation Conference, June 12-16, 2000, Mitzpe Ramon, Israel*, pages 268–271.

SIMON, A., GRAVIER, G. et SÉBILLOT, P. (2013). Leveraging lexical cohesion and disruption for topic segmentation. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1314–1324.

SITBON, L. et BELLOT, P. (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. *In Actes de la conférence Traitement automatique des langues*.

SITBON, L. et BELLOT, P. (2007). Topic segmentation using weighted lexical links (wll). *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07.

STOKES, N. (2003). Spoken and written news story segmentation using lexical chains. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology : Proceedings of the HLT-NAACL 2003 Student Research Workshop - Volume 3*, NAACLstudent '03, pages 49–54.

TÜR, G., HAKKANI-TÜR, D. Z., STOLCKE, A. et SHRIBERG, E. (2001a). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, pages 31–57.

TÜR, G., HAKKANI-TÜR, D. Z., STOLCKE, A. et SHRIBERG, E. (2001b). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

UTIYAMA, M. et ISAHARA, H. (2001). A statistical model for domain-independent text segmentation. *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*.

VITÁNYI, P. M., BALBACH, F. J., CILIBRASI, R. L. et LI, M. (2009). Normalized information distance. *In Information theory and statistical learning*, pages 45–82. Springer.

WAN, S., DRAS, M., PARIS, C. et DALE, R. (2003). Straight to the point : Discovering themes for summary generation. *In Proceedings of the Australasian Language Technology Workshop*.

WANG, X., XIE, L., LU, M., MA, B., CHNG, E. et LI, H. (2012). Broadcast news story segmentation using conditional random fields and multimodal features. *IEICE Transactions*, 95-D(5):1206–1215.

WITBROCK, M. J. et MITTAL, V. O. (1999). Ultra-summarization : A statistical approach to generating highly condensed non-extractive summaries. *In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–316.

XIE, L., YANG, Y., LIU, Z.-Q., FENG, W. et LIU, Z. (2010a). Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using maximum entropy approach. *In International Conference on Audio Language and Image Processing (ICALIP)*, pages 407–413.

XIE, L., YANG, Y., LIU, Z.-Q., FENG, W. et LIU, Z. (2010b). Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using

maximum entropy approach. *In Audio Language and Image Processing (ICALIP)*, pages 407–413.

XIE, L., ZENG, J. et FENG, W. (2008). Multi-scale texttiling for automatic story segmentation in chinese broadcast news. *In AIRS*, pages 345–355.

YAARI, Y. (1998). Texplore- exploring expository texts via hierarchical representation. *In Content Visualization and Intermedia Representations (CVIR'98)*, pages 25–32.

ZHAO, W. X., JIANG, J., HE, J., SONG, Y., ACHANANUPARP, P., LIM, E. et LI, X. (2001). Topical keyphrase extraction from twitter. *In The 49th Annual Meeting of the Association for Computational Linguistics, June, Portland, Oregon, USA*, pages 379–388.

ZHENG, L., LEUNG, C.-C., XIE, L., MA, B. et LI, H. (2012). Acoustic texttiling for story segmentation of spoken documents. *In ICASSP*, pages 5121–5124.