Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Segmentation thématique de transcriptions automatiques de données audiovisuelles

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Lufei LIU

Directeur de mémoire :

Cyril GROUIN

Encadrante :

Camille GUINAUDEAU

Année universitaire 2021/2022

Remerciements

En préambule de ce mémoire, je souhaite adresser mes remerciements à celles et ceux qui m'ont soutenu pour la réalisation de celui-ci.

Je tiens tout d'abord à remercier sincèrement Monsieur Cyril GROUIN, mon encadrant de mémoire, pour m'avoir guidé lors de la réflexion et la rédaction, avec beaucoup de patience et de bienveillance, et en étant toujours disponible lorsque cela s'avérait nécessaire.

Je souhaite ensuite remercier vivement Madame Camille GUINAUDEAU, mon encadrante du stage, pour les nombreux soutiens qu'elle m'a apporté durant le stage ; les échanges et les discussions régulières et intéressantes qui m'ont fait découvrir le côté fascinant de la recherche.

Mes remerciements vont ensuite à l'équipe pédagogique du Master pluriTAL, qui m'a permit de m'aventurer dans le domaine du traitement automatique des langues et qui m'a apporté de l'aide tout au long du Master.

Un grand merci également à mes collègues du LISN pour leur soutien et les discussions quotidiennes sur des sujets de recherche divers et variés.

Enfin, je souhaite remercier ma famille, ainsi que mes amis, qui m'ont toujours encouragé dans les moments difficiles, et qui m'ont poussé à donner le meilleur de moi-même, que ce soit dans la réalisation de ce mémoire ou plus généralement durant toute cette année universitaire.

RÉSUMÉ

Structurer les données audiovisuelles est un travail important au vu de la quantité croissante de celles-ci. L'une des solutions est de les classer en fonction des thèmes abordés, un travail qui pourrait être réalisé, voire automatisé, à travers les techniques de la segmentation thématique automatiques. Segmenter un document en thèmes permet aux utilisateurs d'avoir une idée générale des thèmes traités, et ainsi de cibler plus facilement les informations qui les intéressent. Ce mémoire a pour objectif de comparer plusieurs méthodes de segmentation thématique et d'analyser leur applicabilité pour les transcriptions automatiques de données audiovisuelles. Nous présenterons le fonctionnement des méthodes choisies et les spécificités des données transcrites. Par la suite, nous évaluerons ces méthodes en empruntant un métrique que nous estimons plus objectif pour la tâche.

Mots-clés : segmentation thématique automatique, transcription automatique, thème, cohésion lexicale, métriques d'évaluation.

Table des matières

Remerciements **3**

Résumé **5**

Liste des figures **8**

Liste des tableaux **9**

Introduction **11**

I Contexte général **15**

1 État de l'art **17**
1.1 Introduction 17
1.2 Thème 17
1.3 Segmentation thématique 19
1.4 Métriques d'évaluation 25
1.5 Conclusion 29

II Expérimentations **31**

2 Corpus **33**
2.1 Introduction 33
2.2 Présentation du corpus disponible 33
2.3 Pré-traitement du corpus 40
2.4 Conclusion 45

3 Segmentation thématique **47**
3.1 Introduction 47
3.2 Présentation des méthodes 47
3.3 Paramétrage des méthodes 49
3.4 Segmentation du corpus GMMP-TV 56
3.5 Conclusion 57

4 Perspectives **59**
4.1 Introduction 59
4.2 Améliorer la segmentation thématique 59
4.3 Identification thématique 61
4.4 Conclusion 62

Conclusion générale **63**

Bibliographie **65**

A ANNEXE **69**
A.1 Exemple de bandeau d'informations dans un journal télévisé résumant le sujet en cours de présentation 69
A.2 Un extrait de transcription du journal *Via-occitanie* 69

LISTE DES FIGURES

0.1 Compositon du projet GEM 12
1.1 Illustration de segmentation linéaire (carré rouge) et segmentation hiérarchique (carré jaune) 20
1.2 Un exemple illustrant une frontière proche et une frontière éloignée par rapport à la référence. Les deux hypothèses auront une précision et un rappel de 0. 26
1.3 Un exemple de calcul du P_k avec k = 3. Un rectangle correspond à une unité atomique, un rectangle bleu représente une frontière thématique et une accolade signifie un désaccord. 27
1.4 Un schéma illustrant le calcul de pureté et de couverture 29
2.1 La composition des corpus disponibles carrés bleus : corpus utilisés carrés oranges : corpus non utilisés 34
2.2 Exemple de segment manuellement annoté Ce segment commence à 377.799 secondes et se termine à 551.884 secondes. Il contient deux sous-segments qui commencent respectivement à 377.799 secondes et à 493.607 secondes 35
2.3 Statistiques du corpus W05_15 35
2.4 Statistiques du corpus W07_14 36
2.5 Les statistiques du corpus GMMP-TV Bleu : le nombre de segments par JT Orange : la taille moyenne des segments par JT 37
2.6 Nombre moyen de segments thématiques par JT Seules les émissions inclues dans les trois corpus sont prises en compte bleu : W07_14 ; orange : W05_15 ; gris : GMMP-TV 38
2.7 Taille moyenne des segments thématiques par JT Seules les émissions inclues dans les trois corpus sont prises en compte bleu : W07_14 ; orange : W05_15 ; gris : GMMP-TV 38
2.8 Taille moyenne des groupes de souffle par JT (3 corpus confondus) Pour des raisons de lisibilité, les noms des JT sont remplacés par journal_X. 1-92 : JT du corpus W07_14 93-118 : JT du corpus W05_15 119-128 : JT du corpus GMMP-TV 39
2.9 Deux exemples de transcription et leurs scores de confiance transcription de référence (gauche) : il connaît connaît connaît les légumes il connaît la viande il connaît tout transcription de référence (droite) : et moi je dis à maman mais pourquoi tu nous a abandonnés 41
2.10 Schéma des pré-traitements des corpus 41
2.11 Un exemple de segmentation en locuteurs réalisée par le système LIUM_SpkDiarization colonne 1 : nom du journal télévisé colonne 2 : 1 colonne 3 : id locuteur colonne 4 : temps de début du groupe de souffle colonne 5 : temps de fin du groupe de souffle colonne 6 : caractéristiques du locuteur colonne 7 : transcription automatique des mots prononcés par le locuteur en un groupe de souffle 42
2.12 Un exemple de principe d'alignement GS : groupe de souffle 43

2.13 Exemple d'un extrait de transcription alignée 43
2.14 Annotation manuelle du programme « France3_1920_national » issu du corpus GMMP-TV Chaque balise « » encadre une transcription (les rectangles jaunes identifient une possibilité d'expansion du contenu). . 44
3.1 Le processus de segmentation par Texttiling, Topictiling et Deeptiling . . . 48
3.2 Courbe de similarité entre les blocs de groupes de souffle Gap score : la similarité cosinus entre les groupes de souffle. Sentence Gap Index : numéro de bloc 48
3.3 Exemple d'un extrait de transcription fourni à Texttiling 49
3.4 Exemple d'un extrait de segmentation produit par Texttiling colonne 1 : 1 si le groupe de souffle débute un segment thématique, 0 sinon colonne 2 : temps de début du groupe de souffle colonne 3 : temps de fin du groupe de souffle colonne 4 : le contenu textuel du groupe de souffle 50
3.5 Scores Texttiling avec différentes valeurs de w 50
3.6 Scores Texttiling avec différentes valeurs de k 51
3.7 Les scores de segmentation thématique sur GMMP-TV 57
4.1 L'extrait du programme After foot 60
4.2 Le thème de chaque segment thématique du programme *arte_jt_soir* (hypothèse) 62

Liste des tableaux

2.1 La moyenne et l'écart type du score de confiance par journal télévisé 40
3.1 Les paramètres de Texttiling 50
3.2 Les paramètres de Texttiling 51
3.3 Les différents corpus utilisés pour l'entraînement de modèles thématiques 52
3.4 Les paramètres du Topictiling 53
3.5 Scores Topictiling avec trois modèles thématiques différents 54
3.6 Scores Topictiling avec différentes valeurs de k 54
3.7 Scores Deeptiling avec différentes valeurs de k 55
3.8 Scores obtenus par différents algorithmes de segmentation sur le corpus W05_15 55
3.9 Scores obtenus par différents algorithmes de segmentation sur le corpus GMMP-TV 56
4.1 Le thème de chaque segment thématique du programme *arte_jt_soir* (référence) 61

INTRODUCTION

Présentation générale

Nul ne peut nier qu'aujourd'hui, l'homme est assailli d'informations qu'il se doit de traiter : autrefois seulement transmises via des supports écrits, les informations se diffusent à présent via de multiples canaux, à la radio, la télévision ou encore sur internet. Selon le rapport de l'Institut National de l'Audiovisuel (INA) : 1,132 milliard de vidéos ont été vues sur leur plateforme pour l'année 2020 [1]. Optimiser l'accès et l'utilisation des informations devient impératif sous peine de se retrouver noyé dans le flux de données produites. A ce titre, les progrès technologiques représentent la maladie mais également pourraient aboutir à la découverte du remède. Les techniques de traitement automatique des langues cherchent à proposer des solutions en ce sens, tels que la segmentation thématique automatique, qui consiste en un découpage d'un texte ou d'une séquence de caractères en segments cohérents de telle sorte que les informations soient liées à un même sujet. Cette segmentation propose un moyen de structurer les ressources et ainsi facilite son accès. L'automatisation de celle-ci est nécessaire car une réalisation manuelle est coûteuse, parfois impossible compte tenu de la quantité de données.

La segmentation thématique peut être considérée comme un étape intermédiaire dans d'autres sujets de recherches. Une des applications que l'on peut mentionner est le projet GEM (Gender Equality Monitoring) [2], porté par l'INA, Deezer et des laboratoires STIC et SHS qui est un projet inter-diciplinaire visant à analyser l'égalité des genres dans les médias, au travers un volume considérable de données étalées sur 80 ans.

Le projet est composé de trois axes principaux (Figure 0.1) dont chacun bénéficie des connaissances de domaines différents. Les trois axes représentés sont dépendants les uns des autres. En effet, le premier axe permet la réalisation du second qui serait exploité par le dernier et pourrait permettre d'enrichir les découvertes du premier.

Dans le cadre du projet GEM, la segmentation thématique est un étape nécessaire du fait de la quantité considérable de données à traiter. En effet, afin d'effectuer des analyses plus approfondies sur l'égalité des genres dans les médias, un des aspects à considérer concerne l'identification de sujets abordés dans les émissions télévisuelles et radiophoniques puis, à partir de ces sujets recueillis, d'étudier si oui ou non il existe une corrélation entre le genre et les sujets abordés, comme par exemple si certains sujets sont davantage présentés par les hommes que les femmes ; si les incivilités telles que l'interruption de la parole, les insultes sont plus fréquentes lorsque l'on aborde certains sujets que d'autres et auquel cas, étudier la fréquence de ces incivilités en raison de la présence de femmes ou d'hommes.

1. https://www.ina.fr/sites/default/files/2021-06/RA-2020.pdf
2. Mesure de l'égalité entre les sexes dans les médias – GEM : https://anr.fr/Projet-ANR-19-CE38-0012

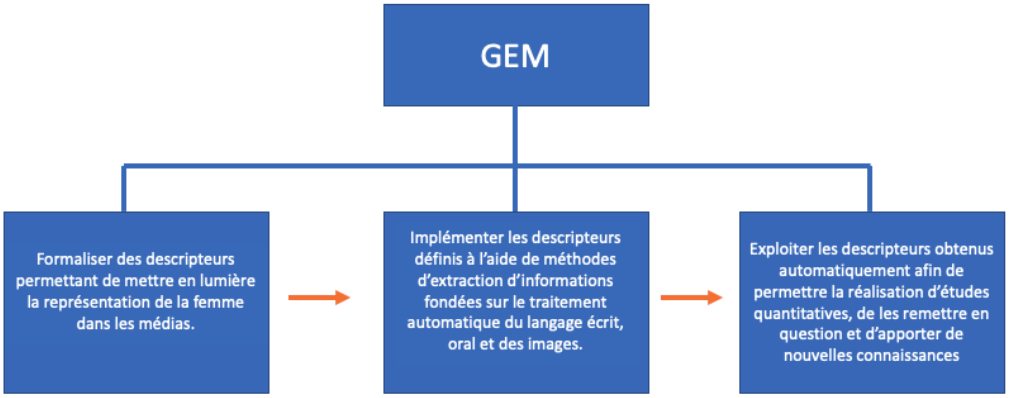


FIGURE 0.1 – Compositon du projet GEM

Problématique et objectif

Les méthodes de segmentation automatique peuvent être classées en deux catégories : elles peuvent soit être supervisées, c'est-à-dire faire apprendre au modèle les caractéristiques de phrases marquant une transition thématique ; ou elles peuvent être non supervisées qui regroupent des phrases du même thème grâce à l'analyse des similarités lexicales. Ces méthodes se révèlent efficaces lorsqu'il s'agit de segmenter des documents écrits. Il est ainsi naturel de penser à les appliquer pour segmenter les documents oraux en utilisant la transcription automatique de ceux-ci. Cependant, il est à noter que les transcriptions des émissions audiovisuelles, en particulier les transcriptions automatiques, se différencient des documents écrits comme des articles ou des journaux sur certains aspects : tout d'abord, les signes de ponctuation sont absents. Deuxièmement, contrairement à un document écrit, généralement découpé en paragraphes représentant chacun une idée principale, un point de vue ou un thème différent, la transcription automatique n'aboutit qu'à un texte sans structure sémantique. Enfin, étant donné qu'aucun système de transcription n'est parfait, les erreurs de transcriptions sont plus ou moins nombreuses, créant ainsi des « mots » qui n'ont pas de sens.

Dans ce travail, nous allons comparer la performance des méthodes de segmentation thématique sur les transcriptions automatiques et tenter d'apporter une réflexion sur les questions suivantes :

1. Les méthodes de segmentation thématique performantes sur des textes écrits sont-elles également applicables aux transcriptions automatiques des émissions audiovisuelles ?

2. Est-il possible d'évaluer une segmentation thématique de manière objective ?

3. L'introduction du modèle de langue pré-entraîné apporte-elle un gain considérable sur la performance de segmentation ?

Plan de lecture

Ce mémoire s'organise en 4 chapitres. Le premier chapitre présentera l'état de l'art en ce qui concerne les méthodes de la segmentation thématique et leur évaluation. Dans un second temps, nous allons œuvrer à présenter en détails les différents corpus qui seront utilisés dans le cadre de nos travaux qui feront l'objet de la partie suivante, à savoir la présentation de plusieurs expériences de segmentation thématique. Enfin, après l'analyse des résultats de ces différences expériences, il sera temps dans la dernière partie de prendre du recul vis à vis de celles-ci et de la segmentation thématique dans son ensemble, afin de possiblement améliorer cette dernière.

Première partie

Contexte général

CHAPITRE 1

ÉTAT DE L'ART

Sommaire

1.1 Introduction **17**
1.2 Thème **17**
1.2.1 Définition 17
1.2.2 Thème dans les données audiovisuelles 18
1.3 Segmentation thématique **19**
1.3.1 Méthodes de segmentation thématique 20
1.3.2 Segmenter les documents oraux 23
1.4 Métriques d'évaluation **25**
1.4.1 Précision, Rappel, F-mesure 25
1.4.2 P_k 26
1.4.3 WindowDiff 27
1.4.4 Pureté, couverture 28
1.5 Conclusion **29**

1.1 Introduction

La segmentation thématique est une tâche courante dans le traitement automatique des langues car elle offre une possibilité de structuration des données permettant une exploitation efficace de celles-ci. Sa réalisation dépend à la fois des théories linguistiques et des progrès en informatique. Dans ce chapitre, nous aborderons en premier lieu la notion de thème dans la littérature ainsi que les nuances apportées à cette notion à partir des documents audiovisuels. Ensuite, nous présenterons les approches développées pour la segmentation thématique, que ce soit pour les documents écrits ou oraux. Enfin, nous expliquerons comment ces approches sont évaluées et nous procéderons à une analyse critique des métriques utilisés.

1.2 Thème

1.2.1 Définition

Qu'est-ce qu'un thème ? Il est difficile de trouver une réponse unique car la notion de "thème" est vaste et non sans ambiguïtés. Si le dictionnaire Larousse propose de définir le "thème" comme un *« sujet, idée sur lesquels portent une réflexion, un discours, une œuvre, autour desquels s'organise une action. »*, les linguistes et les littéraires proposent plusieurs définitions :

- [Collot, 1988] met en évidence une caractéristique récurrente du thème dans une œuvre : *« le thème exprime la relation affective d'un sujet au monde sensible ; il se manifeste dans les textes par une récurrence assortie de variations ; il s'associe à d'autres thèmes pour structurer l'économie sémantique et formelle d'une œuvre. »*.

- [Brémond and Pavel, 1988], en faisant référence aux autres auteurs qui tentent de donner une définition de la notion de thème, avance le rôle « réunissant » du thème dans une œuvre. Lui-même propose de considérer le « thème » comme étant : *« à la rencontre de l'attention référentielle avec l'*à propos de *contenu dans l'œuvre. »*

- [Rastier, 1995] propose de définir le « thème » comme *« une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses. »*

- [Stockinger, 2003] présente le thème comme *« le lieu cognitif servant à capter, à identifier une grandeur donnée comme porteuse d'information (potentielle) pour un agent. »*.

Ces définitions, bien que variées et pouvant sembler complexes, permettent d'extraire une caractéristique commune. Le thème pourrait ainsi être considéré comme le noyau à partir duquel un ensemble d'énoncés se développe. Identifier un thème dans une suite d'énoncés revient donc à trouver "ce dont on parle". Là encore, le travail est loin d'être trivial car celui-ci subit une subjectivité de deux niveaux, qu'elle soit en rapport au contenu du thème ou bien de la granularité :

Pour illustrer la subjectivité du contenu d'un texte, nous pouvons considérer le passage suivant :

Depuis le début de l'été, la France est proie à de fortes vagues de chaleur. Couplées à une sécheresse intense, elles provoquent de forts dégâts dans les champs et élevages de l'Hexagone. Résultat : quelques produits manquent dans les rayons des supermarchés, affectés par ailleurs par la guerre en Ukraine. [1]

Le thème de ce paragraphe, bien que succinct, pourrait ne pas faire l'unanimité si on le présentait à différents lecteurs. En effet, nous pouvons penser qu'il s'agit d'« écologie » si l'on prête davantage l'attention aux « fortes vagues de chaleur » et à la « sécheresse intense ». Néanmoins, si l'on s'intéresse aux termes « manque de produits » ou « affectés par la guerre en Ukraine », on peut aisément penser que le thème mis en évidence est l'« économie ». Cela montre que, pour un même texte, la définition du contenu d'un thème peut varier radicalement selon les personnes.

En ce qui concerne la granularité, celle-ci peut de la même manière faire l'objet de désaccords. En effet, si l'on regarde l'article dans sa globalité, celui-ci se trouverait être les réponses données en matière agricoles aux problèmes soulevés par le réchauffement climatique. Ce thème global pourrait être subdivisé fort logiquement en de nombreux sous-thèmes mais cette subdivision pourrait être toute autre. En effet, un thème général peut englober plusieurs sous-thèmes indépendants les uns des autres mais, en fonction de l'analyse de chacun, les thèmes identifiés pourraient être plus ou moins fins.

1.2.2 Thème dans les données audiovisuelles

Identifier des thèmes n'est pas une tâche exclusive au document écrit. La production croissante de données audiovisuelles rend sa structuration de plus en plus im-

1. Comment réinventer notre agriculture face aux vagues de chaleur et sécheresses? https://www.nouvelobs.com/ecologie/20220822.OBS62251/comment-reinventer-notre-agriculture-face-aux-vagues-de-chaleur-et-secheresses.html

portante. Le concept de thème dans les données audiovisuelles a été abordé dans le cadre de travaux sur les flux multimédias, du point de vue de la détection et du suivi des thématiques abordées dans l'actualité [Allan et al., 1998]. Dans ce projet, la notion de « thème » (topic) se réduit à celle d'« évènement » (event), indiquant « une chose unique qui se produit à un moment donné dans le temps ». Les auteurs explicitent cette définition avec l'exemple suivant : *« l'éruption du Mont Pinatubo le 15 juin 1991 est considérée comme un événement, alors que les éruptions volcaniques en général sont considérées comme une classe d'événements. »*. Le travail de [Guinaudeau, 2011] a pour but d'effectuer une segmentation thématique linéaire des journaux télévisés, un segment thématique étant rapproché à un reportage (éventuellement associé à ses plateaux de lancement et de fin). [Bouchekif, 2016] définit de son côté le thème des journaux télévisés à un niveau relativement fin, en se référant à des découpages thématiques réalisés manuellement par les chaînes de télévision. Un « thème » est ainsi considéré comme une unité non seulement concentrée sur une information précise, qui a lieu à un instant et un endroit donné, mais aussi utilisable dans d'autres tâches automatiques comme la recherche d'informations, le résumé, le titrage automatique, etc.

La difficulté de définir le concept de « thème », que ce soit pour des documents écrits ou oraux, montre à la fois l'ambiguïté de celui-ci mais indique de la même manière la nécessité d'en établir une définition claire et précise en fonction de la tâche dans laquelle elle est impliquée et l'objectif de celle-ci.

L'objectif de ce mémoire est d'utiliser le thème comme critère afin de permettre une segmentation automatique pertinente d'un texte ou plus précisément, de la transcription automatique des données audiovisuelles. Cela passe bien évidemment par la détermination la moins ambiguë possible des sujets abordés dans celui-ci. Nous avons choisi de retenir la définition du thème formulé par [Rastier, 1995]. Nous considérons le thème comme une idée au cœur d'une portion de texte et qui doit s'apprécier par rapport au contenu total du document dans lequel elle apparaît.

1.3 Segmentation thématique

L'objectif de la segmentation thématique est, comme son nom l'indique, de découper un texte en segments cohérents sur le plan thématique. Cette segmentation permet de structurer les informations en fonction des thèmes abordés, ce qui a pour conséquence de faciliter la lecture des dits textes. Par exemple, il est plus simple de lire un article avec des sous-titres précisant le sujet de chaque partie que si celui-ci est présenté en un bloc.

En traitement automatique des langues, la segmentation thématique fait l'objet de nombreuses recherches, que ce soit pour du contenu textuel ou oral. En effet, cette tâche, qui pourrait paraître anodine elle-même, constitue un pré-requis pour de nombreuses applications. Dans la recherche d'informations, la segmentation thématique permet d'obtenir un résultat plus précis, correspondant le plus possible à la requête [Hearst, 1997]. La segmentation thématique constitue également une étape intermédiaire en résumé automatique [Barzilay and Elhadad, 1997, Dias et al., 2007] car elle permet de cibler un thème précis ou de produire un résumé avec des informations plus pertinentes. D'autres tâches telles que l'indexation des documents textuels ou audiovisuels [Amaral and Trancoso, 2003] ou l'analyse de discours

[Galley et al., 2003] pourraient également tirer profit du contenu préalablement segmenté.

1.3.1 Méthodes de segmentation thématique

Etant donnée la complexité de la définition du thème, les méthodes de segmentation thématique se concentrent principalement sur l'identification de changement thématique. Ces méthodes pourraient être globalement divisées en deux branches : la segmentation hiérarchique et la segmentation linéaire. La première, comme son nom indique, cherche à obtenir une structure hiérarchique du document. En effet, un thème peut être abordé sous différents aspects : à partir d'un ou plusieurs thèmes généraux, il est possible de décomposer le document en sous-thèmes dont chacun peut encore être décomposé de manière récursive jusqu'au niveau le plus fin sur le plan de l'unité de sens, qui pourrait être une phrase. La segmentation linéaire découpe quant à elle le document en segments thématiquement cohérents tout en restant sur le niveau le plus général. Après la segmentation, les sujets des segments différents sont totalement indépendants. La figure 1.1 illustre la différence entre ces deux types de segmentations.

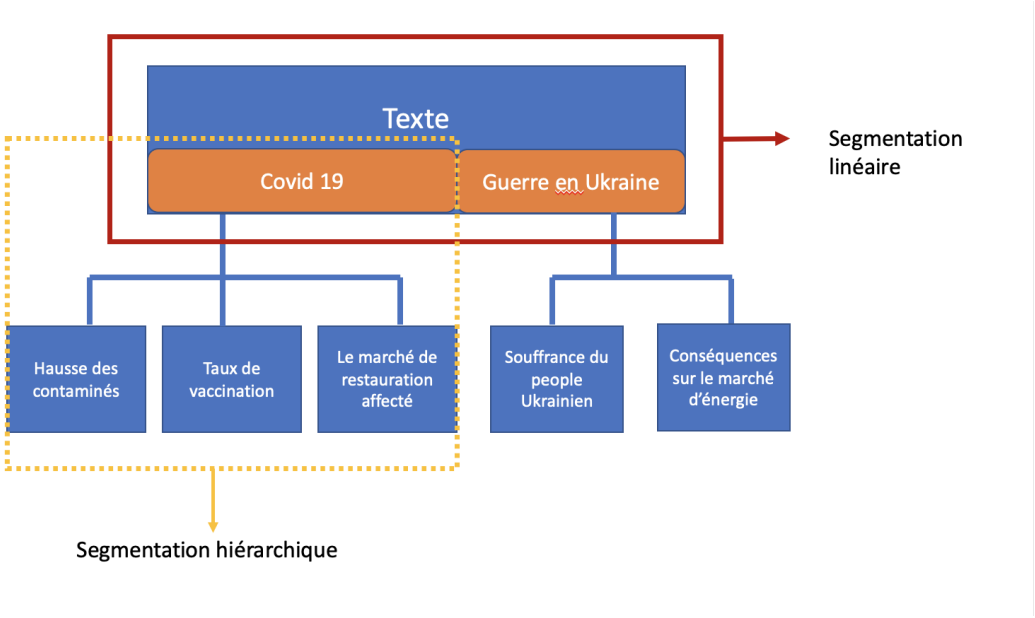


FIGURE 1.1 – Illustration de segmentation linéaire (carré rouge) et segmentation hiérarchique (carré jaune)

Si la notion de structure hiérarchique semble évidente pour un humain, identifier cette structure paraît plus complexe tant pour un humain que pour la machine. En effet, plus on va dans le détail sur une partie du texte, plus la frontière entre les champs sémantiques des différentes portions de celui-ci est floue. On peut faire une analogie avec la distinction entre plusieurs espèces animales : tandis qu'il est très simple de distinguer un chien d'un canard, parmi les différentes race de chiens, il peut s'avérer être difficile de distinguer un carlin d'un bouledogue français. Ainsi, s'il est difficile pour l'humain de se mettre d'accord sur le niveau de hiérarchie d'un texte, cela se trouve être encore plus difficile pour la machine (et pour l'évaluation du résultat produit par cette dernière). De plus, la segmentation thématique est très

souvent considérée comme une tâche intermédiaire. Il n'est pas toujours nécessaire d'avancer à un niveau très fin. Nous avons opté pour une segmentation linéaire dans notre travail car celle-ci correspond plus à nos besoins.

En ce qui concerne les méthodes de segmentation, celles-ci peuvent varier en fonction des données à disposition. Dans la partie suivante, nous allons passer en revue les techniques principales, qui sont regroupées dans des méthodes statistiques ou des méthodes neuronales.

Méthodes statistiques

Initialement, les techniques de segmentation thématique ont été appliquées sur des documents écrits. Une des méthodes couramment utilisées est celle basée sur la cohésion lexicale afin d'identifier les transitions thématiques du texte. Cette cohésion crée et assure une structure cohérente du texte, permettant de tirer un fil conducteur des idées au sein de celui-ci. Elle se manifeste par la répétition de certains mots, l'utilisation de synonymes et des co-occurrences sémantiquement liées [Halliday and Hasan, 1976, Morris and Hirst, 1991]. On peut ainsi supposer que le vocabulaire peut être similaire à l'intérieur d'un segment thématique, mais être radicalement différent lorsqu'on passe d'un segment thématique à un autre.

Parmi les algorithmes exploitant cette cohésion, l'algorithme Texttiling [Hearst, 1997] est l'une des méthodes fondatrices dont l'idée est reprise dans de nombreux travaux de ce domaine. Cette méthode consiste à parcourir le texte de manière linéaire et de détecter les ruptures de cohésion lexicale à l'aide d'une fenêtre glissante, à travers un calcul de similarité. Précisément, la méthode comprend principalement les étapes suivantes :

1. Découper le texte en pseudo-phrases dont la taille w est pré-définie. Ceci a pour objectif d'assurer une homogénéité entre la taille de chaque phrase à comparer afin d'éviter de pénaliser les phrases courtes par rapport aux phrases longues dans le calcul.

2. Regrouper les phrases en blocs de taille k.

3. Construire une représentation vectorielle de chaque bloc de texte. Il s'agit d'une représentation du type sac de mots dont la taille de vecteur correspond à la taille du vocabulaire (sans mots-outils et lemmatisé) du texte. La valeur de chaque élément du vecteur correspond à la fréquence du mot dans le bloc.

4. Calculer la similarité cosinus entre les blocs adjacents. Les valeurs obtenues permettent de tracer une courbe de la cohésion lexicale dont les pics indiquent une forte ressemblance lexicale et les vallées représentent une faible similarité lexicale.

5. Identifier les frontières thématiques à l'aide d'un calcul de score de profondeur. Ces scores mesurent la distance entre les pics à chaque côté d'une vallée. Ainsi plus la vallée est profonde, plus le score est élevé. Enfin, ces scores sont comparés à un seuil minimal dans le but de déterminer si la vallée correspond ou non à une frontière thématique.

La fréquence des mots est également utilisée dans l'algorithme C99 [Choi, 2000] pour mesurer la cohésion lexicale. Contrairement à Texttiling, la similarité est calculée entre chaque paire de phrases pour construire une matrice de similarité. Ensuite, une opération de classement est effectuée sur chaque élément de la matrice avant de procéder à l'identification de segments thématiques à l'aide d'une méthode de clustering basée sur l'algorithme de maximisation [Reynar, 1994].

[Galley et al., 2003] développe LCseg qui reprend l'idée de Texttiling mais utilise la chaîne lexicale pour identifier les répétitions des termes. La similarité entre les

blocs de texte est ensuite calculée à partir des valeurs des chaînes pondérées avec la méthode term frequency-inverse document frequency (TF-IDF) [Sparck Jones, 1972].

Ces méthodes basées sur la similarité lexicale ont prouvé leurs performances mais ne sont pas exemptes de défauts, notamment :

- L'utilisation de synonymes et la co-référence n'est pas prise en compte. En effet, d'un point de vue stylistique, la réutilisation excessive d'un même terme n'est pas idéale. On peut ainsi trouver des phrases traitant de sujets similaires mais dont le vocabulaire est différent sur la forme, alors qu'ils sont pourtant sémantiquement liés sur le fond.

- Le nombre élevé de dimensions dans les vecteurs des mots. Le vocabulaire d'un texte étant souvent très large, cela peut conduire à un vecteur de mots de taille considérable, ce qui pourrait ralentir le calcul et n'est pas forcément représentatif d'informations.

Segmentation à l'aide des modèles thématiques Avec l'introduction des modèles thématiques (topic model) en TAL, les méthodes de segmentation thématique se voient également évoluer dans ce sens. Plusieurs auteurs se penchent sur le modèle de l'Allocation Dirichlet Latente (LDA) [Blei et al., 2001] qui est un modèle génératif probabiliste capable d'inférer les thèmes abordés dans un document. Ces méthodes reposant sur la modélisation thématique ont contribué à pallier le problème de sparsité des vecteurs. Nous trouvons par exemple le travail de [Sun et al., 2008] qui utilise le modèle LDA pour identifier les thèmes latents dans les blocs composés d'une séquence de phrases. Un noyau Fisher (Fisher Kernel) est ensuite utilisé pour calculer la similarité entre les blocs adjacents. La frontière thématique est établie via une programmation dynamique en fonction du score obtenu par chaque bloc, basé sur sa longueur et sa similarité sémantique avec son précédent. [Misra et al., 2009] ont adapté l'algorithme U00 [Utiyama and Isahara, 2001] en intégrant les informations thématiques obtenues par LDA afin d'effectuer non seulement la segmentation mais aussi l'étiquetage thématique des segments. De plus, leur modèle est d'abord entraîné sur une grande quantité de données avant d'être utilisé pour segmenter des textes que le modèle n'a jamais vu auparavant. [Riedl and Biemann, 2012b] proposent Topictiling qui optimise Texttiling sur la représentation du texte. L'algorithme, au lieu de calculer la similarité entre deux vecteurs de mots, mesure celle entre deux vecteurs de thèmes : supposons qu'un modèle LDA avec T thèmes est obtenu avec des données d'entraînement. Ce modèle annote chaque mot du texte à segmenter en leur attribuant un identifiant thématique. Ensuite, chaque bloc (composé de k phrases) est représenté par un vecteur de taille T dont chaque élément correspond à la fréquence du thème dans le bloc. L'algorithme Topictiling permet de créer des vecteurs bien plus compacts, en plus de permettre un calcul de manière linéaire par rapport aux nombres de phrases, ce qui rend sa complexité moins importante que ceux qui utilisent la programmation dynamique.

Méthodes neuronales

L'arrivée de l'apprentissage profond a révolutionné le TAL avec des architectures neuronales complexes et puissantes. Naturellement, cette technique est également utilisée pour développer des approches de segmentation thématique. [Koshorek et al., 2018] ont développé une approche neuronale supervisée qui prédit pour chaque phrase la probabilité qu'elle mette fin à un segment thématique. Le modèle est composé de deux sous-réseaux hiérarchisés de type long short-term

memory (LSTM) [Hochreiter and Schmidhuber, 1997]. Le réseau du niveau inférieur comprend deux couches d'architecture LSTM bidirectionnelles qui génèrent un plongement pour chaque phrase à partir des mots qui les composent. Ces plongements alimentent ensuite le réseau du niveau supérieur pour effectuer la prédiction. [Glavaš and Somasundaran, 2020] proposent Coherence-Aware Text Segmentation (CATS), un modèle neuronal similaire à celui de [Koshorek et al., 2018] mais qui apporte deux originalités : premièrement, le modèle utilise deux niveaux de réseaux Transformer [Vaswani et al., 2017] pour encoder le texte : le premier produit une représentation vectorielle de chaque phrase à partir des plongements des tokens et de leurs positions tandis que le deuxième prend en entrée la sortie du premier et produit une nouvelle représentation pour chaque phrase en intégrant les informations dans leurs contextes. Deuxièmement, en plus de la prédiction de segmentation au niveau de phrases, le modèle effectue également une prédiction auxiliaire de cohérence des segments. En effet, le modèle est entraîné pour pouvoir distinguer des vrais segments cohérents à des segments crées de manière artificielle avec des phrases n'ayant aucun lien entre elles. Cette prédiction, couplée à la prédiction de segmentation, permet au modèle d'atteindre une meilleure performance par rapport aux nombreuses autres méthodes neuronales.

Nous pouvons constater que les modèles utilisés deviennent plus complexes mais permettent également de prendre en compte de plus en plus d'informations, en particulier celles liées au contexte.

1.3.2 Segmenter les documents oraux

Le développement technologique a pour conséquence que l'écrit n'est plus l'unique moyen de véhiculer des informations. Avec l'essor des nouvelles technologies, de nombreuses formes de données audiovisuelles ont fleuri, que ce soient les contenus informatifs tels que les journaux télévisés et radiophoniques ou les contenus divertissants comme les séries télévisées, les téléréalités, etc. S'ajoute à cela les vlogs [2] démocratisés par les réseaux sociaux ; les plate-formes de diffusion de flux rendent accessible à tous des activités nécessitant auparavant une présence physique, faisant apparaître davantage de ressources audiovisuelles avec, entre autres, les conférences monologues (TED talk), ou bien encore l'enregistrement en ligne de séminaires.

Si les textes écrits ont souvent pour caractéristique une structure, claire et concise [3], cela est moins le cas des documents oraux, tout du moins pour certains types de contenu comme les interviews, les débats télévisés dont la parole est souvent spontanée et la durée est relativement importante. L'intérêt de structurer ces données en thèmes se révèle être primordiale car cela permet de faciliter l'accès à l'information recherchée et aboutit à une recherche plus efficace sans y passer trop de temps. Par exemple, on peut mentionner les vidéos de conférence TED qui, par la mention des thèmes dans les timecodes à des moments précis permettent de ne visualiser seulement la partie qui nous intéresse. On peut également penser que la segmentation de la transcription d'une réunion en thèmes peut permettre aux personnes concernées de cibler une information plus rapidement.

2. vlog ou blog vidéo, est un type de blog dont le principal média est la vidéo, pouvant être commentée ou non par ses visiteurs. https://fr.wikipedia.org/wiki/Vlog

3. Les contenus sur les réseaux sociaux font une exception mais ici par textes écrits, nous nous concentrons sur des documents soigneusement produits comme un article scientifique, un journal écrit, etc.

Méthodes basées sur la transcription

Les méthodes basées sur la cohésion lexicale ont montré leurs performances sur la segmentation des textes écrits, et leurs applications sur la segmentation des émissions audiovisuelles est également envisagée avec comme source la transcription de celles-ci. La transcription manuelle étant une tâche coûteuse, surtout face à l'augmentation rapide des données orales, les techniques de transcription automatique ont largement contribué à faciliter l'accès à celle-ci. [Georgescul et al., 2006] a obtenu des résultats assez satisfaisants en utilisant Texttiling sur la transcription automatique des réunions. [Galley et al., 2003] proposent de considérer la segmentation thématique comme une tâche de classification binaire. Cela revient ainsi à décider si une phrase représente une frontière thématique. Ils suggèrent d'utiliser des expressions de repère (cue phrases) comme l'une des caractéristiques de classification. En effet, des travaux précédant démontrent que certains particules de discours pourraient être porteuses d'informations sur la structure du texte. Ainsi, les auteurs sélectionnent des expressions récurrentes dans les phrases débutant des segments thématiques et les intègrent dans les critères de classification. Ces méthodes transformant la segmentation des documents oraux en la segmentation des documents écrits ont obtenu de résultats encourageants. Cependant, leurs sources principales sont issues des techniques de transcription automatique qui certes, s'améliorent de manière continue, mais qui restent tout de même imparfaites. La présence d'erreurs de transcriptions pourrait directement impacter la performance de la segmentation si cette dernière repose complètement sur les indices lexicaux. [Guinaudeau et al., 2012] proposent ainsi d'exploiter les mesures de confiance pour ajuster le poids de chaque mot automatiquement transcrit dans le calcul de cohésion lexicale. En effet, pour chaque mot transcrit un score de confiance lui est accordé indiquant la probabilité que celui-ci soit correctement transcrit. De plus, la relation sémantique entre les mots est intégrée dans le calcul de cohésion afin de surmonter les problèmes liés aux faibles répétitions lexicales. Ainsi, les mots correctement transcrits contribuent davantage que ceux qui ne le sont pas. L'intégration de relations sémantiques est également étudiée dans d'autres travaux : [Nie et al., 2013] étudient la similarité sémantique entre chaque paire de mot en prenant en compte leur contexte. Cette similarité est ensuite propagée sur l'ensemble du vocabulaire et est intégrée dans le calcul de similarité entre les phrases. [Bouchekif et al., 2015] améliorent la segmentation thématique des journaux télévisés en intégrant dans le calcul de cohésion les relations sémantiques extraites à partir des journaux écrits produits à la même période que les journaux télévisés. Pour éviter d'entraîner un modèle sur des données externes, [Ghinassi, 2021] introduit l'utilisation des encodeurs neuronaux de phrases (Neural Sentence Encoders) dans son algorithme Deeptiling. L'algorithme repose sur le principe développé dans Texttiling mais utilise des encodeurs neuronaux pour produire une représentation vectorielle riche en informations pour chaque phrase des documents à segmenter. L'utilisation d'un encodeur multilingue permet même au modèle d'être robuste face à des langues différentes. [Sheikh et al., 2017] innovent sur la manière de mesurer la cohésion lexicale avec une architecture de réseaux de neurones récurrents (RNN) bidirectionnels composés des cellules LSTM. Le modèle détecte le changement thématique en comparant la similarité des contextes capturés à partir des séquences de mots avant et après le mot courant. L'évaluation du modèle est effectuée à la fois sur la transcription des émissions d'actualités concaténées mais aussi sur celle des programmes d'actualités réelles. En revanche, le modèle s'est avéré plus performant sur les données artificielles. Cela mérite d'ailleurs notre attention car

l'évaluation sur un jeu de données synthétiques ne peut pas refléter la performance réelle du modèle de manière objective.

Méthodes multimodales

La transcription représente une source essentielle à la segmentation des documents oraux. Cependant, les documents multimédias contiennent des indices divers et variés qui leurs sont propres et pourraient également être porteurs d'informations thématiques. Certains chercheurs tentent d'exploiter, voire de combiner ces indices de différentes modalités afin de produire une segmentation plus satisfaisante. [Bouchekif et al., 2014] supposent qu'il existe non seulement une cohésion lexicale mais aussi une cohésion des locuteurs au sein d'un segment thématique, en particulier pour les journaux télévisés. Si les présentateurs restent tout au long de l'émission, certains journalistes ou invités n'interviennent que sur un thème donné. Il est ainsi pertinent d'intégrer la distribution des locuteurs dans le calcul de similarité. Dans les journaux télévisés, le passage d'un sujet à un autre est souvent marqué par un retour au plateau du présentateur ou la présence de publicités, une identification automatique de ces indices visuels [Hmayda et al., 2020, Hauptmann and Witbrock, 1998] pourrait être combinée aux autres indices utiles à la segmentation comme la détection des silences [Galley et al., 2003]. Une approche multimodale a été mise en avant par [Dumont and Quénot, 2012], intégrant les indices visuels et acoustiques tels que le visage des présentateurs, les images des logos, les textes affichés à l'écran, le silence, etc. Il est à noter que la combinaison des indices pourrait certes être bénéfique à la segmentation, mais certains d'entre eux pourraient être spécifiques aux données, ce qui rend la méthode sensible aux différents styles des émissions. Par exemple, le visage de présentateurs pourrait être un indice significatif pour la segmentation des journaux télévisés mais le serait bien moins pour segmenter une conférence monologue telle qu'une conférence TED. Le choix des indices multimodaux est ainsi crucial à la performance des méthodes de cette famille.

1.4 Métriques d'évaluation

L'évaluation est une étape non négligeable dans le traitement automatique des langues, cependant sa réalisation n'est pas toujours évidente. La segmentation thématique étant une tâche relativement subjective, son évaluation en est de même, tant sur le plan de la production de références que sur le plan des métriques. Évaluer l'efficacité des algorithmes de segmentation revient principalement à mesurer l'accord entre une segmentation hypothétique et celle de référence. Nous présenterons par la suite quelques métriques couramment utilisés pour la segmentation thématique et nous introduirons également une paire de métriques : pureté et couverture que nous estimons adaptés à notre évaluation.

1.4.1 Précision, Rappel, F-mesure

La précision, le rappel et la f-mesure ont été définis pour la recherche d'information. Ces mesures ont ensuite été utilisées pour évaluer d'autres tâches, y compris la segmentation thématique. Pour cette dernière, le calcul revient à quantifier les frontières proposées et s'effectue de la manière suivante :

$$\text{Précision} = \frac{\text{nombre de frontières correctement proposées}}{\text{nombre total de frontières proposées}} \tag{1.1}$$

$$\text{Rappel} = \frac{\text{nombre total de frontières correctement proposées}}{\text{nombre total de frontières attendues}} \tag{1.2}$$

$$\text{F-mesure} = \frac{(1+\beta^2) \times \text{ précision } \times \text{ rappel}}{\beta^2 \times \text{ précision } + \text{ rappel}} \tag{1.3}$$

La F-mesure (formule 1.3) désigne la moyenne harmonique pondérée du rappel et de la précision. La valeur accordée à β permet de donner plus de poids au rappel ou à la précision, ou encore d'équilibrer les deux mesures (avec $\beta = 1$).

La précision et le rappel évaluent la présence ou non d'une frontière à sa place de référence : une frontière hypothétique est considérée comme correcte si et seulement si celle-ci est placée exactement au même endroit que celle de référence. Ainsi, une frontière hypothétique au voisinage de celle de référence est pénalisée de la même manière qu'une autre beaucoup plus éloignée, alors que ces deux erreurs ne sont pas du tout au même niveau(figure 1.2). Une telle mesure est bien trop sévère pour évaluer la segmentation thématique et ne peut pas refléter la performance des algorithmes de manière juste. Bien que la F-mesure permet d'ajuster l'importance accordée à la précision ou au rappel, le résultat n'est guère interprétable. [Guinaudeau et al., 2012] propose par ailleurs d'instaurer une tolérance pour pallier au problème de la mise en correspondance exacte des frontières, mais cette tolérance reste une valeur subjective.

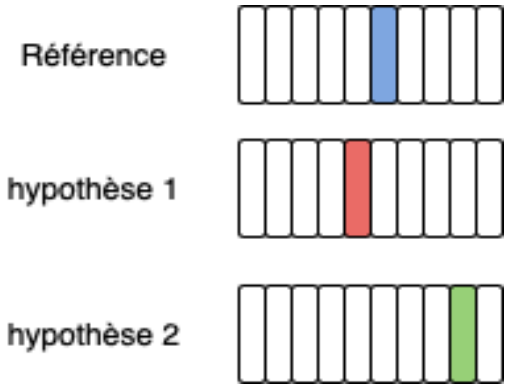


FIGURE 1.2 – Un exemple illustrant une frontière proche et une frontière éloignée par rapport à la référence.
Les deux hypothèses auront une précision et un rappel de 0.

1.4.2 P_k

Le taux d'erreur P_k est proposé par [Beeferman et al., 1999] pour tenir compte des problématiques présentées ci-dessus. Le calcul du P_k repose sur une fenêtre glissante de taille k correspondant au nombre d'unités atomiques (tokens, phrases, groupes de souffle, etc.) à prendre en compte. Les auteurs ont suggéré une valeur optimale du paramètre k, à savoir la moitié de la taille moyenne des segments dans la segmentation de référence.

En faisant glisser la fenêtre de manière parallèle sur la segmentation de référence R et la segmentation hypothétique H proposée par l'outil à évaluer, le métrique Pk

regarde si les deux extrémités de la fenêtre se trouvent dans le même segment, à la fois dans la référence et dans l'hypothèse, et produit un désaccord si cela n'est pas le cas. Le calcul revient donc à compter le nombre de fois où le désaccord se produit. La formule de calcul est la suivante :

N : Le nombre total d'unités atomiques dans le texte ;

r_i : la $i^{ème}$ unité atomique de la segmentation de référence ;

h_i : la $i^{ème}$ unité atomique de la segmentation d'hypothèse ;

$$P_k(R,H) = \frac{1}{\text{N - k}} \sum_{i=1}^{N-k} f(f(r_i, r_{i+k}), f(h_i, h_{i+k})) \tag{1.4}$$

On peut illustrer le calcul avec l'exemple suivant (Figure 1.3) :

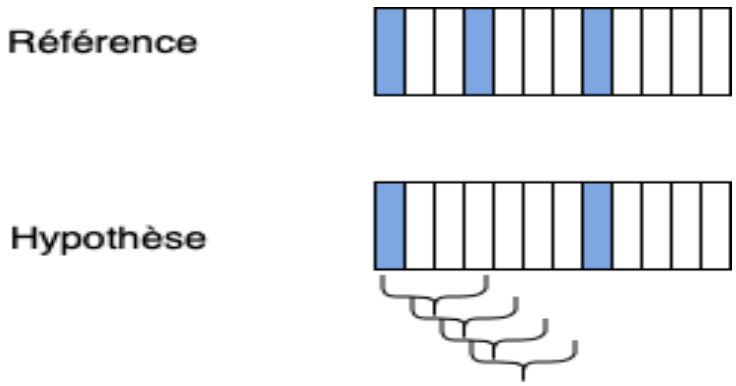


FIGURE 1.3 – Un exemple de calcul du P_k avec k = 3. Un rectangle correspond à une unité atomique, un rectangle bleu représente une frontière thématique et une accolade signifie un désaccord.

Dans cet exemple, la segmentation est effectuée sur un texte composé de 12 phrases (supposons qu'une phrase représente une unité atomique) avec une fenêtre de taille 3. Nous pouvons constater l'omission d'une frontière dans l'hypothèse, ce qui entraîne 4 désaccords avec la référence. Ainsi, le taux d'erreur P_k = $\frac{4}{12-3} \approx 0.44$.

1.4.3 WindowDiff

Le métrique P_k est certes optimisé mais n'est pas sans failles. Celles-ci ont été pointées par [Pevzner and Hearst, 2002] : premièrement, la mesure est très sensible à la taille de fenêtre. Dans l'exemple ci-dessus, le nombre de désaccords ne serait pas identique si le paramètre k était fixé à une autre valeur. Deuxièmement, une sous-segmentation (faux négatif) est davantage pénalisée qu'une sur-segmentation (faux positif). Enfin, les frontières hypothétiques incorrectes mais proches de la référence sont trop pénalisées.

En conséquence, les auteurs ont apporté une modification à P_k en proposant la mesure WindowDiff qui fonctionne de la manière suivante : pour chaque fenêtre de taille k, il suffit de comparer le nombre de frontières de référence qui s'y trouvent (r_i) avec le nombre de frontières hypothétiques (h_i). L'algorithme est pénalisé si $r_i \neq h_i$. Formellement, le score WindowDiff est calculé ainsi :

N : Le nombre total d'unités atomiques dans le texte ;

b(r_i,r_{i+k}) : le nombre de frontières entre la $i^{ème}$ et la $(i+k)^{ème}$ unité atomique de la segmentation de référence ;

b(h_i,h_{i+k}) : le nombre de frontières entre la $i^{ème}$ et la $(i+k)^{ème}$ unité atomique de la segmentation proposée par l'algorithme ;

$$\text{WindowDiff}(R,H) = \frac{1}{\text{N - k}} \sum_{i=1}^{N-k} (|b(r_i, r_{i+k}) - b(h_i, h_{i+k}| > 0) \tag{1.5}$$

Selon les auteurs, WindowDiff permet de traiter de manière équitable les sous-segmentations et les sur-segmentations. De plus, il détecte les frontières incorrectes aussi bien pour de petits segments que pour de grands segments.

Bien que la mesure WindowDiff se soit imposée comme un indice de référence dans de nombreux travaux, elle est toutefois sujette à de nombreuses critiques. En effet, elle pénalise de manière différente les erreurs en fonction de leur position dans le texte [Lamprier et al., 2007]. De plus, elle reste tout de même sensible à la variation de la taille de segments même si celle-ci est plus réduite par rapport à P_k.

1.4.4 Pureté, couverture

Les métriques P_k et WindowDiff, bien que populaires, présentent deux principaux défauts : d'abord, ils sont relativement sensibles à la variation de taille des segments. Ensuite, ils se focalisent sur les frontières posées et non sur les segments dans leur ensemble, rendant ainsi les valeurs numériques produites difficilement interprétables. Compte tenu de ces problèmes, nous proposons une paire de métriques : pureté et couverture qui sont couramment utilisés pour évaluer la détection automatique des locuteurs.

$\max_{référence} |hypothèse \cap référence|$: L'intersection entre la durée de chaque paire de segment de référence et d'hypothèse. Si plusieurs segments d'hypothèses chevauchent un segment de référence, celui dont la co-occurrence est la plus élevée sera pris en compte.

$$\text{Pureté} = \frac{\sum_{hypothèse} \max_{référence} |hypothèse \cap référence|}{\sum_{hypothèse} |hypothèse|} \tag{1.6}$$

$$\text{Couverture} = \frac{\sum_{référence} \max_{hypothèse} |référence \cap hypothèse|}{\sum_{référence} |référence|} \tag{1.7}$$

Le figure 1.4 fournit un exemple de calcul de pureté et de couverture. Chaque segment est délimité par leur timecode, c'est-à-dire le temps de début et de fin. Le calcul du ratio est donc basé sur la durée : la couverture mesure le rapport entre référence et hypothèse, et inversement pour la pureté. Dans l'exemple ci-dessous, la couverture du segment 1 dans la référence est de 100% car la totalité de celui-ci est couvert par le segment A de l'hypothèse. En revanche, la couverture du segment 3 dans la référence est de 90% car 90% de celui-ci est couvert par le segment C dans l'hypothèse. Bien que le segment B couvre aussi une partie de segment 3, cette part est largement inférieure à celle du segment C, ainsi seul le segment 3 est comparé au segment C. Le raisonnement est similaire en ce qui concerne le calcul de la pureté.

La paire de métriques pureté et couverture mesure la qualité de la segmentation en comparant les paires de segments de référence et d'hypothèse dans leur ensemble, au lieu de simplement prendre en compte la position ou les nombres de frontières

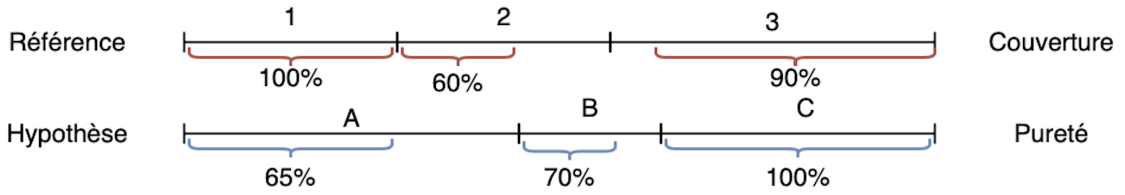


FIGURE 1.4 – Un schéma illustrant le calcul de pureté et de couverture

retrouvées. Cette évaluation permet de contourner la nécessité de définir une taille de fenêtre au préalable, dont le changement de valeur peut avoir une incidence sur le résultat de l'évaluation. Nous obtenons ainsi un résultat constant avec ces métriques.

1.5 Conclusion

La segmentation thématique est une tâche nécessaire mais complexe. Bien que les techniques de ce champs d'études soient multiples et relativement robustes, elles ne sont pas adaptées à tout types de données, comme notamment à la segmentation thématique des données audiovisuelles du fait de la présence d'erreurs de transcriptions et à cause du manque de données annotées. Outre l'étude de diverses méthodes, l'évaluation de celles-ci mérite également notre attention car elle demeure complexe. En effet, la production du corpus de référence peut se révéler être subjective, en particulier lorsqu'il s'agit d'obtenir un accord inter-annotateurs. De plus, les métriques de l'évaluation sont de leur côté loin d'être parfaits. C'est pourquoi dans le cadre de notre travail, nous avons emprunté les métriques utilisés pour la détection automatique des locuteurs, à savoir la pureté et la couverture, ce qui permet d'évaluer les méthodes sans être dépendant des éléments variables comme la taille du fenêtre.

Deuxième partie

Expérimentations

CHAPITRE 2

CORPUS

Sommaire

2.1 Introduction **33**
2.2 Présentation du corpus disponible **33**
2.2.1 FrNewsLink 34
2.2.2 GMMP-TV 36
2.2.3 Analyse du corpus 37
2.3 Pré-traitement du corpus **40**
2.3.1 Alignement du corpus FrNewsLink 42
2.3.2 Annotation du corpus GMMP-TV 43
2.4 Conclusion **45**

2.1 Introduction

Dans ce chapitre, nous allons présenter les corpus utilisés pour effectuer la comparaison de différents modèles capables d'identifier les changements thématiques d'un texte. Dans la section 2.2, nous détaillerons la composition du corpus. Dans la section 2.3, nous présenterons les pré-traitements effectués sur ces corpus à des fins variées.

2.2 Présentation du corpus disponible

Dans ce travail, nous avons à notre disposition trois corpus comme indiqué sur la figure 2.1. Il est à noter que tous les corpus ne sont pas utilisés dans nos expériences. En effet, certains corpus n'étant pas annotés, il n'y a pas de référence pour évaluer les algorithmes. De plus, les corpus non utilisés sont composés d'émissions radiophoniques dont la structure thématique est assez différente d'un journal télévisé, en particulier le corpus2 qui contient les libres antennes, les magazines d'informations, les émissions musicales et les interviews. Il est possible qu'un algorithme performant sur les journaux télévisés ne le soit pas sur les émissions radiophoniques. Compte tenu de ces limites de corpus, nous avons choisi de concentrer notre travail sur les journaux télévisés.

Dans cette partie, nous allons présenter en détail les corpus utilisés pour nos expériences, à savoir le corpus FrNewsLink et le corpus GMMP-TV. Le premier a principalement servi comme corpus de paramétrage et d'entraînement tandis que le second est utilisé comme corpus de test.

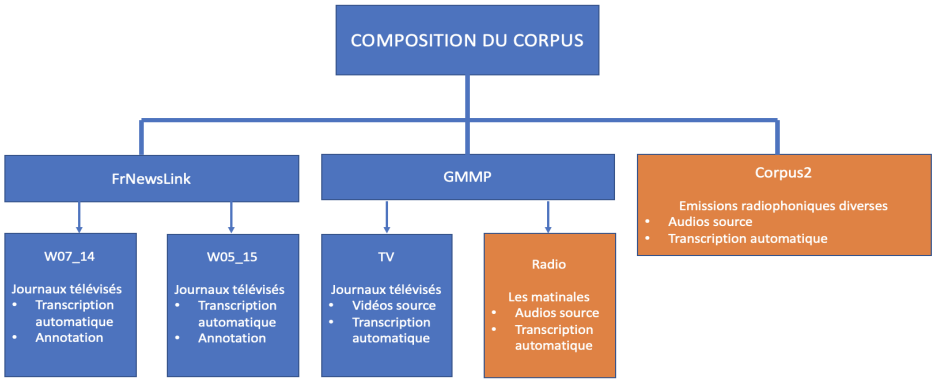


FIGURE 2.1 – La composition des corpus disponibles
carrés bleus : corpus utilisés
carrés oranges : corpus non utilisés

2.2.1 FrNewsLink

Le corpus FrNewsLink[1] est construit par [Camelin et al., 2018] et contient les transcriptions de journaux télévisés ainsi que de presses écrites collectées sur internet. Il est composé de deux sous-corpus de transcriptions :

- le corpus W07_14 : comprend les transcriptions de 92 journaux télévisés de 8 chaînes pour la septième semaine de l'année 2014 et représentant une durée totale d'environ 23 heures ;

- le corpus W05_15 : ce corpus inclut les transcriptions de 26 journaux télévisées des 26 et 27 janvier 2015, pour une durée totale de 9,9 heures.

- un corpus de presses écrites : 22 141 articles de presses issus de la page principale de Google News ont été collectés sur les mêmes périodes, ayant pour objectif d'associer le contenu de ceux-ci à ceux des journaux télévisés. Ce corpus écrit n'est pas utilisé dans notre travail car nous nous concentrons sur la transcription automatique des émissions.

En raison des droits d'utilisation, les ressources audiovisuelles sont absentes du corpus. Cependant, le contenu de la transcription automatique, de la segmentation en locuteurs (en anglais *speaker diarization*)[2] ainsi que l'annotation sont présents. La transcription est effectuée avec le système LIUM ASR et la segmentation en locuteurs est réalisée avec le système LIUM_SpkDiarization [Meignier and Merlin, 2010]. En ce qui concerne l'annotation, chaque journal télévisé est découpé manuellement aux endroits où un changement thématique a lieu. Ainsi, pour chaque journal est produit un fichier d'annotation dans lequel est inscrit le temps de début et de fin de chaque segment thématique ainsi qu'une description succinte du thème abordé. Dans le cas où un segment peut encore être divisé en plusieurs sous-segments, une description de chacun sera également fournie avec les indications de temps de début. La figure 2.2 illustre un exemple de segment annoté.

Etant donné que nous travaillons sur la segmentation thématique, il serait intéressant de connaître le nombre de segments thématiques par programme et la

1. https://lium.univ-lemans.fr/frnewslink/
2. La segmentation en locuteur consiste en la partition d'un flux sonore en segments selon l'identité du locuteur. L'objectif est de savoir « qui parle à quel moment ».

```
<Section type="report" topic="to8" startTime="377.799" endTime="551.884">
<Turn startTime="377.799" endTime="551.884">→ temps début et fin du segment
<Sync time="377.799"/>→ temps début du sous-segment 1
Intempéries en Grande-Bretagne : le Sud-Ouest est sous les eaux et s'apprête à faire face à une
nouvelle tempête→ sous-segment 1
<Sync time="493.607"/> → temps début du sous-segment 2
Le premier ministre David Cameron demande de l'aide à l'Union Européenne →sous-segment 2
</Turn>
</Section>
```

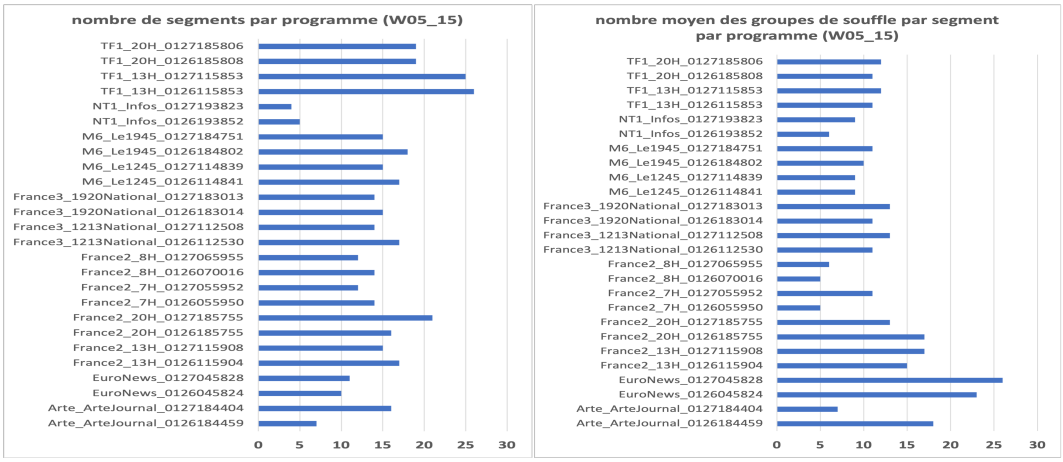
FIGURE 2.2 – Exemple de segment manuellement annoté
Ce segment commence à 377.799 secondes et se termine à 551.884 secondes.
Il contient deux sous-segments qui commencent respectivement à 377.799 secondes et à 493.607 secondes

taille des segments. De manière générale, la taille d'un segment thématique peut être mesurée par le nombre d'unités minimales (token, phrase, etc.) dans celui-ci. Etant donné qu'il n'existe pas de ponctuation dans la parole, nous travaillons à l'échelle du groupe de souffle, défini comme un ensemble de syllabes prononcées en un souffle continu. Ainsi, pour les deux sous corpus du FrNewsLink, nous avons calculé :

- le nombre de segments : le nombre de thèmes identifiés par l'annotateur pour chaque journal télévisé.

- la taille moyenne des segments : le nombre moyen de groupes de souffle par segment pour chaque journal télévisé.

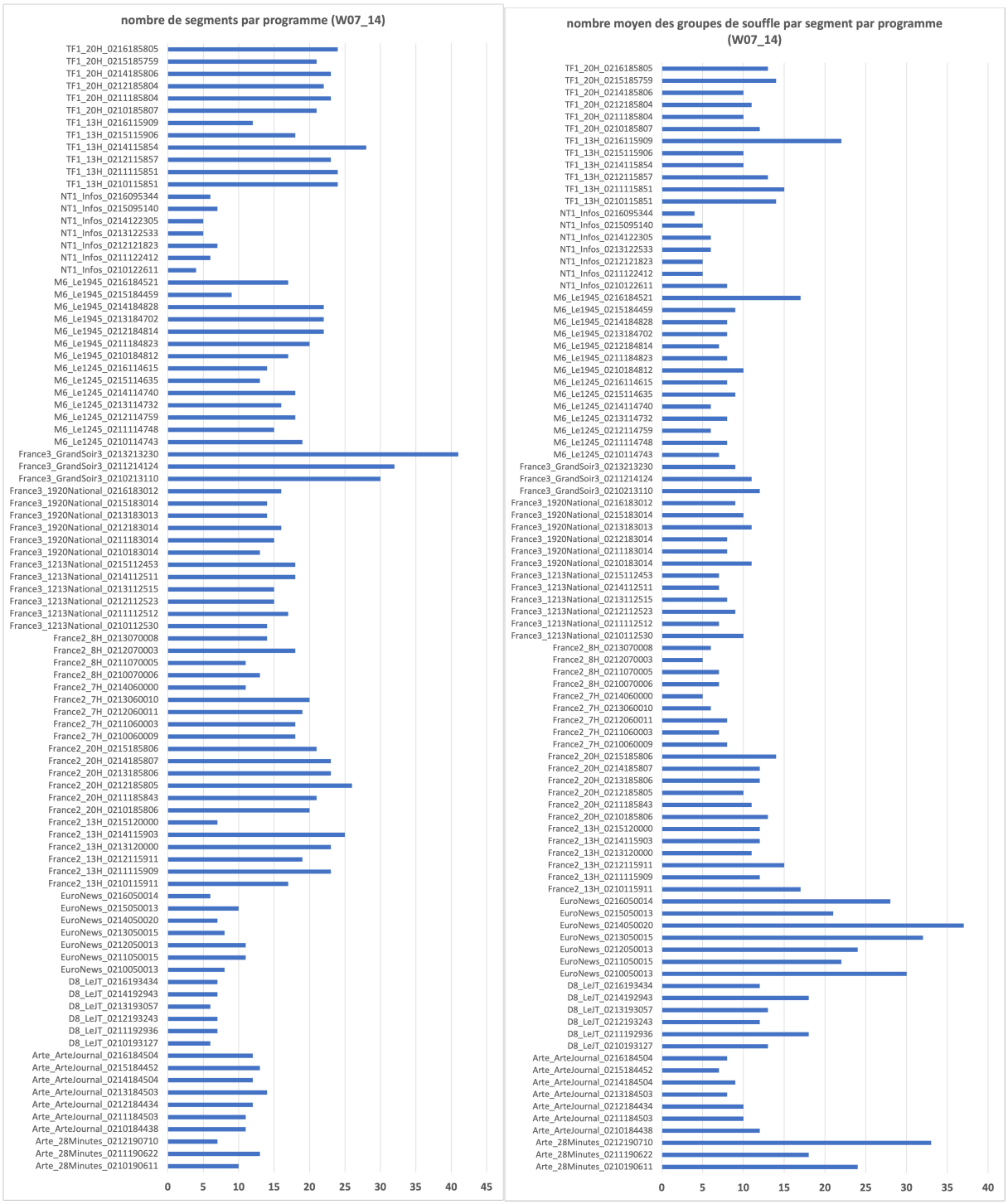
Sur la figure 2.3(a), nous pouvons observer que pour le corpus W05_15, le nombre de segments thématiques pour une même émission est similaire quel que soit le jour de diffusion du programme. En ce qui concerne la taille de segments qui varie en fonction de la richesse de chaque thème abordé, nous avons calculé, pour chaque programme du corpus, la taille moyenne des segments. La figure 2.3(b) montre que les programmes d'une même émission contiennent des segments de taille plus ou moins similaire. Ces caractéristiques sont également relevées dans le corpus W07_14 (2.4(a), 2.4(b)).



(a) Nombre de segments thématiques par JT : W05_15

(b) Taille moyenne des segments thématiques par JT : W05_15

FIGURE 2.3 – Statistiques du corpus W05_15



(a) nombre de segments thématiques par JT : W07_14 (b) La taille moyenne des segments thématiques par JT : W07_14

FIGURE 2.4 – Statistiques du corpus W07_14

2.2.2 GMMP-TV

Le corpus GMMP-TV est composé de 10 journaux télévisés du 29 septembre 2020, collectés pour étudier la présence des femmes dans les médias dans le cadre du « Glo-

bal Media Monitoring Project »[Coulomb-Gully, 2012]. L'ensemble des programmes a été transcrit automatiquement avec le système LIUM ASR. Par ailleurs, chaque programme est accompagné d'un fichier CSV décrivant la nature de chaque groupe de souffle, s'il s'agit d'un locuteur homme ou femme, la présence de musique, etc. Il est à noter que les journaux télévisés contenus dans ce corpus ne sont pas annotés. Afin de créer un corpus de référence permettant d'évaluer les modèles de segmentation thématique, nous avons procédé à une annotation manuelle consistant à indiquer le changement de thème à partir de la transcription automatique. Ce travail d'annotation sera détaillé dans la section suivante (section 2.3.2).

Pour ce corpus, nous avons également calculé pour chaque épisode, le nombre de segments thématiques ainsi que la taille moyenne de ceux-ci (figure 2.5) en nous basant sur l'annotation produite.

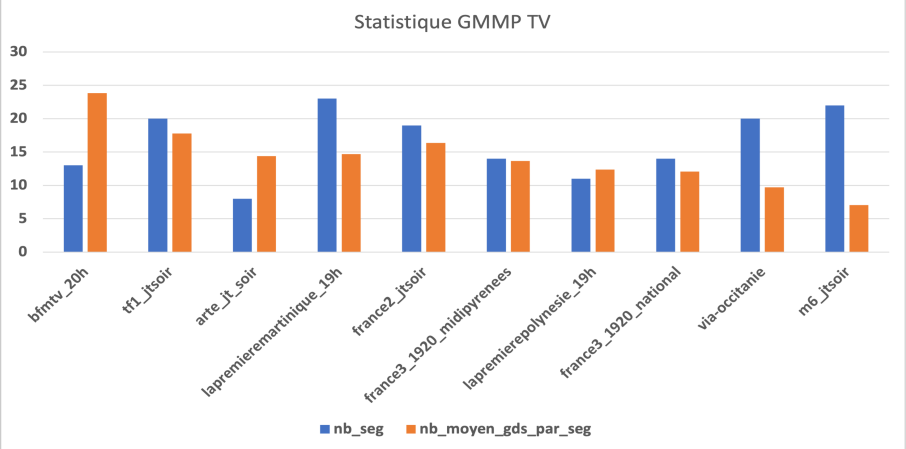


FIGURE 2.5 – Les statistiques du corpus GMMP-TV
Bleu : le nombre de segments par JT
Orange : la taille moyenne des segments par JT

2.2.3 Analyse du corpus

Comparaison avec FrNewsLink Si ces trois corpus ont été choisis pour nos expériences, c'est parce qu'ils se distinguent les uns des autres mais présentent également certaines similarités.

Tout d'abord, en termes de type d'émission, quasiment toutes les émissions [3] présentes dans le corpus W05_15 se trouvent également dans le corpus W07_14, à l'exception de Arte_28 minutes, D8_LeJT, et France3_GrandSoir3. Parmi les dix émissions du corpus GMMP-TV, la moitié est présente dans les deux corpus précédents. Deuxièmement, pour les mêmes émissions, le nombre de segments thématiques (figure 2.6) ainsi que la taille moyenne des segments (figure 2.7) restent similaires, bien qu'ils soient diffusés à des périodes différentes. Finalement, la taille moyenne des groupes de souffle, c'est-à-dire le nombre de tokens [4] compris dans un groupe de souffle se situe principalement dans le même intervalle pour les trois corpus : entre 25 et 35 (figure 2.8).

3. On distingue « émission » au « programme »/« épisode ». Une émission désigne un journal, par exemple TF1_13H, mais un programme/épisode désigne un journal télévisé d'une date précise.

4. Les tokens de chaque JT transcrit sont obtenus à partir d'un processus de tokenization réalisée avec spaCy.

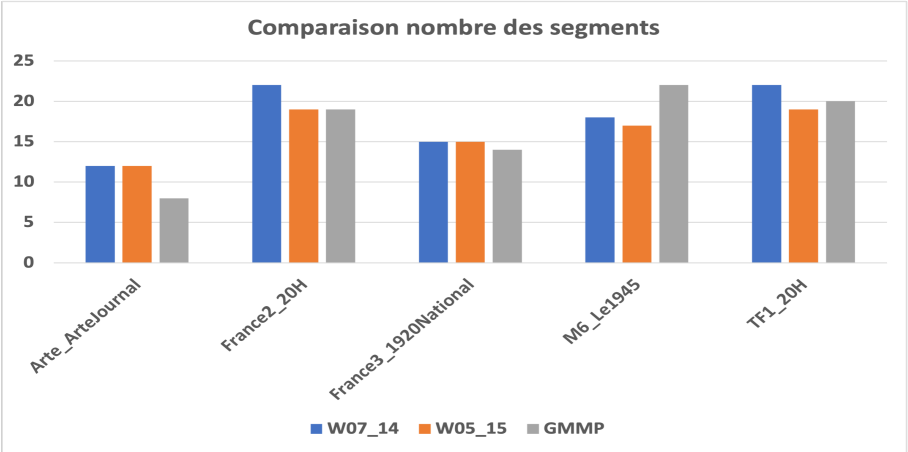


FIGURE 2.6 – Nombre moyen de segments thématiques par JT
Seules les émissions inclues dans les trois corpus sont prises en compte
bleu : W07_14 ; orange : W05_15 ; gris : GMMP-TV

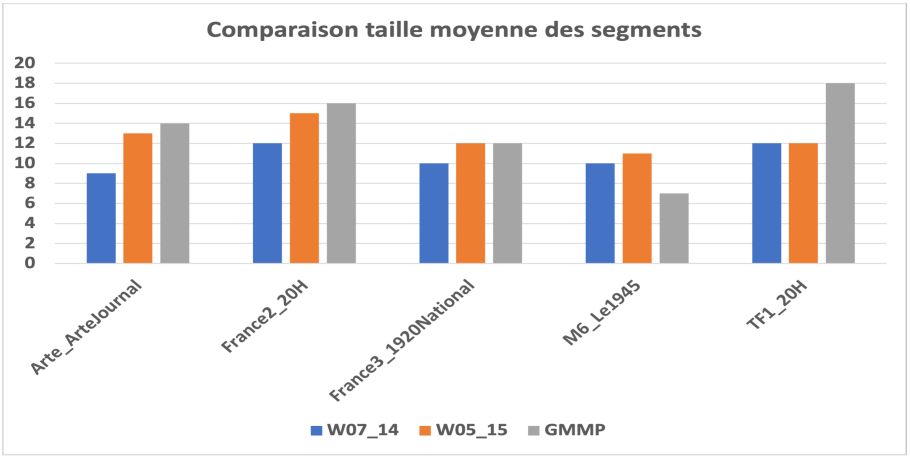


FIGURE 2.7 – Taille moyenne des segments thématiques par JT
Seules les émissions inclues dans les trois corpus sont prises en compte
bleu : W07_14 ; orange : W05_15 ; gris : GMMP-TV

Hormis les similarités mentionnées ci-dessus, les trois corpus se distinguent les uns des autres sur les points suivants :

- la diversité des émissions : les trois corpus partagent certaines émissions mais chacun rassemble également des émissions qui n'existent pas dans les autres ;

- le nombre d'échantillons : pour chacun des trois corpus, le nombre d'épisodes recueillis pour chaque émission varie. Dans W07_14, ce nombre se situe entre trois et sept, deux pour W05_15 et un seul pour GMMP-TV ;

- les périodes de diffusions : les journaux télévisés dans W07_14, W05_15 et GMMP-TV sont respectivement diffusés en février 2014, janvier 2015 et septembre 2020. A priori, les actualités ne se répètent pas d'une année à l'autre, le contenu de chaque corpus est ainsi assez éloigné.

Compte tenu des caractéristiques et de la répartition de contenus des trois corpus, nous avons décidé d'utiliser le corpus W07_14 comme corpus d'entraînement et de paramétrage, le corpus W05_15 comme corpus de développement afin de réajuster les paramètres en cas de besoin et le corpus GMMP-TV comme corpus de test afin

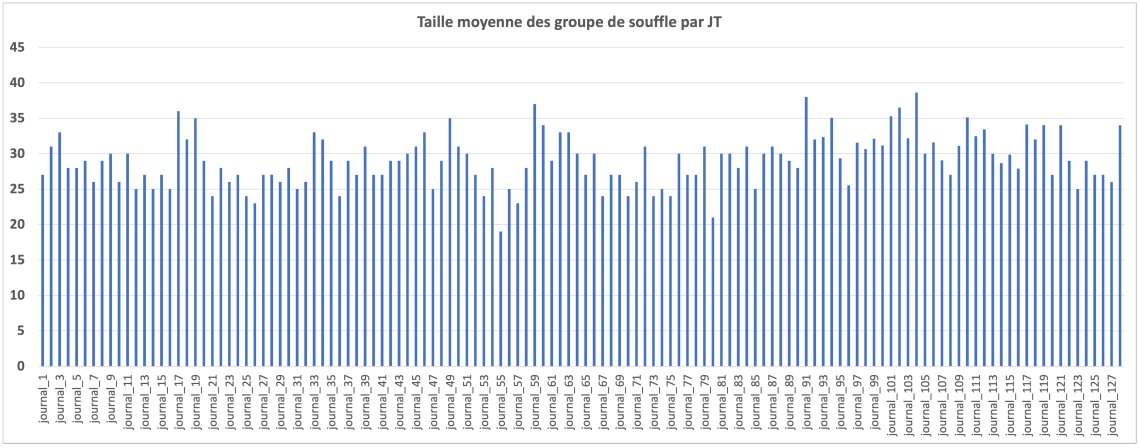


FIGURE 2.8 – Taille moyenne des groupes de souffle par JT (3 corpus confondus)
Pour des raisons de lisibilité, les noms des JT sont remplacés par journal_X.
1-92 : JT du corpus W07_14
93-118 : JT du corpus W05_15
119-128 : JT du corpus GMMP-TV

d'évaluer les algorithmes à comparer.

Erreurs de transcription Comme nous l'avons mentionné précédemment, le corpus textuel à notre disposition est la transcription automatique réalisée par le système LIUM ASR. Ce système produit un taux d'erreurs de mots (WER pour *word error rate*)[5] entre 9% et 13% sur un corpus similaire au corpus FrNews-Link [Camelin et al., 2018]. Afin d'avoir une idée qualitative de la transcription du corpus GMMP-TV, nous avons d'abord pensé à nous baser sur le score de confiance obtenu par chaque mot transcrit. Ce score, compris entre 0 et 1 est produit par le système de reconnaissance automatique de la parole afin d'estimer la probabilité qu'un mot soit correctement reconnu par le système.

Nous avons effectué une moyenne du score de confiance pour chaque groupe de souffle, puis avons effectué une moyenne pour l'ensemble des groupes de souffle d'un journal télévisé (table 2.1). Nous pouvons constater que pour chaque journal, le score moyen est assez élevé. Afin de vérifier s'il existe des disparités sur un journal donné, nous avons décidé également d'effectuer une analyse de l'écart type. Les valeurs obtenues sont assez proches et faibles, ce qui montre une moindre disparité entre les scores de confiance. Cependant, il est à noter que l'écart type est plus élevé pour le journal *Via-occitanie*, qui représente également la plus faible moyenne. En effet, ce journal d'environ 33 minutes contient 9 minutes de publicités qui ne sont presque pas transcrites, ce qui produit des scores assez bas (annexe A.2).

Cependant, les scores de confiance sont-ils réellement le miroir de la qualité de transcription ? Pour répondre à cette question, nous avons choisi d'analyser quelques groupes de souffle dont le score moyen est inférieur à 0.8. Nous constatons qu'il peut arriver que les mots correctement transcrits reçoivent un score assez faible ou que les transcriptions erronées obtiennent une score plutôt élevé. Cette situation arrive surtout lors de passages accompagnés d'un fond de musique assez fort, ou avec un locuteur qui a un accent assez prononcé.

5. WER mesure la distance d'édition entre une transcription de référence et d'hypothèse. Plus le taux est faible (minimum 0.0) plus la transcription est bonne.

nom_JT	moyenne	écart type
tf1_jtsoir	0.9	0.076
arte_jt_soir	0.918	0.07
bfmtv_20h	0.89	0.095
lapremierepolynesie_19h	0.868	0.104
m6_jtsoir	0.893	0.098
france3_1920_national	0.917	0.067
lapremieremartinique_19h	0.893	0.085
via-occitanie	0.825	0.16
france3_1920_midipyrenees	0.907	0.079
france2_jtsoir	0.899	0.095

TABLE 2.1 – La moyenne et l'écart type du score de confiance par journal télévisé

Nous montrons ici les scores et la transcription de deux groupes de souffle comme exemple (figure 2.9). Au vu du temps dont nous disposons, nous n'avons pas pu regarder en détail chaque groupe de souffle, mais cette découverte nous permet de prendre du recul vis-à-vis des scores obtenus pour l'analyse des erreurs de transcription.

Nous avons également observé quelques autres particularités de transcription que nous estimons nécessaires de mentionner :

1. On peut noter qu'un même terme peut être transcrit de différentes manières au sein du même journal. Par exemple les noms « Joe Biden » et « Donald Trump » peuvent être correctement transcrits dans certains cas, mais dans d'autres cas sont devenus « giovanni », « job haidon », « donald train ». Ces erreurs apparaissent surtout au début du journal pendant l'annonce des titres.
2. On remarque également que le terme covid est mal transcrit (koweit, coville, kovi, vie dix-neuf, etc.) , et ce dans tous les 10 journaux. Cela pourrait traduire une certaine difficulté à transcrire un terme dont l'usage est récent.

Toutes ces erreurs de transcriptions peuvent avoir des conséquences sur la compréhension d'un texte. Pour reprendre l'exemple du terme covid, celui-ci a parfois été transcrit en Koweit : on peut donc légitimement penser qu'il y a un problème dans ce pays, alors que ce n'était pas le sujet.

2.3 Pré-traitement du corpus

Dans cette section, nous allons présenter les pré-traitements que nous avons réalisés. Ces pré-traitements ne sont pas identiques d'un corpus à l'autre en raison des différences de format d'origine et des algorithmes de segmentation utilisés par la suite. En général, l'ensemble de documents des trois corpus se compose de deux versions : une version au format texte brut et une version lemmatisée au format texte brut.

0.87	quand		0.78	et
0.87	il		1.00	moi
0.76	connaît		0.56	je
0.84	qu'		0.54	dis
0.87	on		0.51	à
0.65	est		0.74	ma
0.88	qu'		0.67	ma
0.89	on		0.95	mais
0.57	est		1.00	pourquoi
1.00	les		0.46	tu
1.00	légumes		0.80	nous
0.93	et		0.84	a
0.68	connaît		0.58	abandonnés
0.99	la			
0.95	viande			
0.68	et			
0.66	qu'			
0.70	on			
0.65	est			
0.41	au			

FIGURE 2.9 – Deux exemples de transcription et leurs scores de confiance
transcription de référence (gauche) : il connaît connaît connaît les légumes il connaît la viande il connaît tout
transcription de référence (droite) : et moi je dis à maman mais pourquoi tu nous a abandonnés

La figure 2.10 décrit le schéma des pré-traitements effectués sur les corpus afin de créer un jeu de données de référence et un jeu de données d'entraînement.

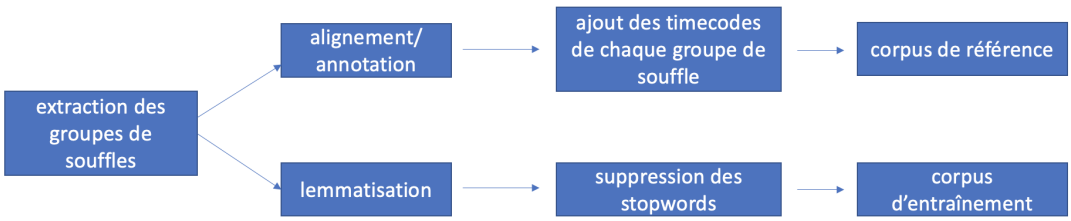


FIGURE 2.10 – Schéma des pré-traitements des corpus

L'extraction des groupes de souffle est réalisée à partir des fichiers de transcription. Ensuite, pour le corpus FrNewsLink, un travail d'alignement est effectué entre l'annotation et les textes extraits afin de faciliter la visualisation de la frontière thématique au sein du texte. Ensuite, nous avons ajouté des informations temporelles de chaque groupe de souffle dans le fichier aligné de chaque émission car celles-ci sont nécessaires à l'évaluation des algorithmes de segmentation automatique.

Afin de construire un jeu de données d'entraînement adapté aux algorithmes uti-

lisés, nous avons, après avoir extrait les textes, appliqué à ceux-ci la lemmatisation et la suppression des mots-outils. Ces deux étapes sont effectuées avec l'outil spaCy [6]. Cependant, il est à noter que la sortie de ces deux étapes n'est pas parfaite compte tenu de la nature des textes en entrée, à savoir la transcription automatique. En effet, spaCy utilise des règles pour trouver le lemme de chacune des formes. Ainsi si certaines formes ne sont pas correctement transcrites ou ne correspondent à aucune règle, alors elles ne pourraient donc ne pas être correctement lemmatisées. De manière similaire, la suppression des mots outils s'effectue en confrontant les textes à une liste des mots : ainsi, sa qualité peut également être impactée par celle de la transcription.

2.3.1 Alignement du corpus FrNewsLink

Comme nous l'avons mentionné précédemment, une annotation manuelle a été effectuée sur le corpus FrNewsLink permettant de repérer les changements thématiques au sein de chaque journal télévisé. Cependant, l'annotation indique seulement l'intervalle de temps de chaque segment thématique, sans contenir le contenu textuel de celui-ci (figure 2.2). Un travail d'alignement est ainsi réalisé pour retrouver les transcriptions correspondant à chaque segment thématique. Cela permet de faciliter à la fois l'évaluation automatique de l'algorithme et la comparaison manuelle entre la segmentation de référence et celle produite de manière automatique.

Pour effectuer l'alignement, nous avons utilisé les transcriptions organisées selon la segmentation en locuteurs (figure 2.11) car le groupe de souffle est également l'unité minimale pour cette segmentation.

```
1  Arte_28Minutes_2014-02-10-19-06-11 1 S0 0.000 2.140 <o,f0,female> assurance gestion de patrimoine
2  Arte_28Minutes_2014-02-10-19-06-11 1 S2 8.310 10.070 <o,f0,male>
3  Arte_28Minutes_2014-02-10-19-06-11 1 S10 13.900 33.000 <o,f0,female> bonsoir et bienvenue dans vingt minutes au
   sommaire le sport et la politique ou plutôt le sport peut il être politique une question qui nous taraude à la vue
   des jo de sotchi véritable plate-forme de propagande nationaliste et expressions de lubrice d' un autocrate
   vladimir poutine le sport peut -il se soustraire à la politique on va en débattre tout à l' heure et trois athlètes
   de la pensée
4  Arte_28Minutes_2014-02-10-19-06-11 1 S10 33.000 44.150 <o,f0,female> mais tout d' abord je reçois philippe torreton
   un athlète des planches et pas que des planches héros expérimental d' un projet bee media d' arte une enquête
   policière sur l' antenne et un procès à vivre sur le web bonsoir philippe torreton
```

FIGURE 2.11 – Un exemple de segmentation en locuteurs réalisée par le système LIUM_SpkDiarization
colonne 1 : nom du journal télévisé
colonne 2 : 1
colonne 3 : id locuteur
colonne 4 : temps de début du groupe de souffle
colonne 5 : temps de fin du groupe de souffle
colonne 6 : caractéristiques du locuteur
colonne 7 : transcription automatique des mots prononcés par le locuteur en un groupe de souffle

Etant donné que nous nous concentrons sur les thèmes à gros grain, l'alignement se base ainsi sur les attributs « startTime » et « endTime » de chaque balise « <Section> ». Il s'agit donc de rassembler un ensemble de groupes de souffle dont les temps de début et les temps de fin sont inclus dans l'intervalle temporel d'un segment thématique. Cependant, l'annotation manuelle n'étant pas faite sur les groupes de souffle, le début ou la fin d'un groupe de souffle automatiquement transcrits ne

6. https ://spacy.io

correspond pas toujours à celui d'un segment thématique (figure 2.12). Afin de faciliter l'alignement, nous avons décidé d'inclure le groupe de souffle dont le début ou la fin dépasse le segment correspondant dans celui-ci.

FIGURE 2.12 – Un exemple de principe d'alignement
GS : groupe de souffle

Une fois aligné, pour chaque groupe de souffle débutant un segment, l'étiquette "1" lui sera attribué et "0" pour les autres groupes de souffle du même segment (figure 2.13). Parfois, une marque "####COUPURE####" est présente derrière un groupe de souffle, il s'agit des chevauchements de temps mentionnés ci-dessus. Ces marqueurs permettent de faire des statistiques de ce cas particulier si nécessaire.

```
1  1   assurance gestion de patrimoine
2  1
3  1   bonsoir et bienvenue dans vingt minutes au sommaire le sport et la politique ou plutôt le sport peut
   il être politique une question qui nous taraude à la vue des jo de sotchi véritable plate-forme de
   propagande nationaliste et expressions de lubrice d' un autocrate vladimir poutine le sport peut -il se
   soustraire à la politique on va en débattre tout à l' heure et trois athlètes de la pensée  ####COUPURE####
4  0   mais tout d' abord je reçois philippe torreton un athlète des planches et pas que des planches héros
   expérimental d' un projet bee media d' arte une enquête policière sur l' antenne et un procès à vivre sur
   le web bonsoir philippe torreton
5  0   bonsoir je présente nadia damon chaque et vincent giret bonsoir intime conviction et c' est un
   événement donc wait une chaîne pionnière
6  0   mais c' est c' est formidable de de d' avoir eu cette audace la commune et de donner les moyens de
   parler de vous parlez gênant de la chaîne en effet une euh
```

FIGURE 2.13 – Exemple d'un extrait de transcription alignée

2.3.2 Annotation du corpus GMMP-TV

Comme nous l'avons mentionné précédemment, il est nécessaire de produire un corpus de référence pour évaluer les algorithmes de segmentation. Nous avons donc décidé de repérer manuellement les frontières thématiques des dix journaux télévisés du corpus GMMP_TV. Le « thème » étant un concept soumis à diverses interprétations (section 1.2), nous avons choisi de définir un « thème » comme « une idée étant le cœur d'une portion de texte et qui est indépendante de celle qui la précède et de celle qui suit ». Par exemple, le thème « Covid » peut rassembler tous les sujets provenant de journaux d'informations traitant de la hausse du nombre des cas, des difficultés d'approvisionnement en masque ou des mesures de lutte contre la pandémie. L'annotation est produite au format XML (figure 2.14), avec des balises « » marquant les frontières de chaque thème et les attributs « <topic> » fournissant un résumé de chaque thème. Le contenu textuel de chaque thème est compris entre les balises « <text> » à l'intérieur de chaque paire de balise « ».

Les journaux télévisés sont relativement bien structurés, néanmoins, certaines difficultés peuvent survenir durant l'annotation manuelle. Tout d'abord, il est diffi-

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE document SYSTEM "annot-gem.dtd">
<document annotator="Lufei LIU" audio_filename="france3_1920_national" version="1" version_date="210722"/>
<segments>
    <segment topic="annonce des titres">
        <text>
            bonsoir et bienvenue dans ce dix neuf vingt à la une de l' actualité de ce mardi
            villefontaine sous le choc après la mort de victorine que s' est il passé après
            son dernier coup de téléphone à sa famille samedi.
            un appel à témoins lancé nous ferons le point sur l' enquête le gouvernement
            débloque un virgule quatre milliard d' euros pour les salariés des epad en première
            ligne depuis le début de la crise du co vide alors que la situation redevienne
            très préoccupant les maisons de retraite peinent à retrouver.
            et malgré l' accélération de la pandémie en europe certains croisiéristes reprennent
            et traversé mais avec un maximum de précautions sanitaire l' une de nos équipes a
            embarqué en méditerranée.
            enfin le coup d' envoi de la cinq g en france les enchères pour l' attribution
            des fréquences ont débuté aujourd'hui dans un climat de défiance impact sur la
            santé sur l' environnement et sur le porte monnaie le très haut débit.
            va -t-il nous coûter plus cher réponse.
            avec david bowie.
        </text>
    </segment>
    <segment topic="disparition de victorine"> ...
    </segment>
    <segment topic="manque de personnel aux ephads"> ...
    </segment>
    <segment topic="exénoration des cotisations sociales pour les restaurateurs"> ...
    </segment>
    <segment topic="comment les croisières s'adaptent en temps de l'épidémie"> ...
    </segment>
    <segment topic="attaque à proximité des anciens locaux de charlie hebdo"> ...
    </segment>
    <segment topic="l'insécurité à bordeaux"> ...
    </segment>
    <segment topic="tension en biélorussie"> ...
    </segment>
    <segment topic="élection présidentielle américaine"> ...
    </segment>
    <segment topic="des effets de 5g sur la santé"> ...
    </segment>
    <segment topic="bonus auto les primes revues à la baisse"> ...
    </segment>
    <segment topic="fin des animaux aux cirques"> ...
    </segment>
    <segment topic="innovation secouriste"> ...
    </segment>
    <segment topic="fin du journal"> ...
    </segment>
</segments>
```

FIGURE 2.14 – Annotation manuelle du programme « France3_1920_national » issu du corpus GMMP-TV
Chaque balise « » encadre une transcription (les rectangles jaunes identifient une possibilité d'expansion du contenu).

cile de se baser uniquement sur la transcription pour annoter car l'absence de ponctuations et la présence des erreurs de transcription impactent la compréhension du contenu. Deuxièmement, pour les émissions contenant une grande partie de débats, il peut être parfois difficile d'identifier clairement les frontières thématiques. Cela est principalement dû au fait que lors de ceux-ci les participants ont tendance à s'interrompre mutuellement, ajoutant de la cacophonie et donc entravant la bonne lisibilité des différents sujets. Compte tenu de ces difficultés, l'annotation manuelle est d'avantage appuyée sur trois types d'indices, à la fois textuels et visuels :

1. L'annonce des titres au début (parfois au milieu) de chaque journal. Celle-ci donne une idée générale des sujets qui seront traités par la suite dont chacun correspond potentiellement à un segment thématique.
2. Un indicateur visuel permettant de délimiter le début ou la fin d'un thème peut être le retour sur le plateau : en effet, lorsque l'on voit à nouveau le présentateur

ou la présentatrice, cela implique dans la majorité des cas, la transition entre un sujet et un autre.

3. Un autre indicateur visuel peut être lié aux différents bandeaux affichés en bas de l'écran (annexe A.1) : en effet, celui-ci ayant pour but de décrire de manière succincte le contenu du sujet abordé, si l'on se rend compte que celui-ci se trouve être modifié, cela doit en théorie s'accompagner d'un changement thématique, et donc d'un nouveau segment à délimiter.

2.4 Conclusion

Nous avons, dans cette partie, présenté les corpus à notre disposition ainsi que les pré-traitements réalisés sur ceux-ci. Nos corpus étant constitués de transcriptions automatiques d'émissions audiovisuelles et radiophoniques, les éventuelles erreurs de transcriptions pourraient porter atteinte à la performance des algorithmes de segmentation thématique. Afin de comparer l'adaptabilité des différents méthodes à ce type de contenu, nous allons effectuer des expériences de segmentation avec ces corpus, ce qui fera l'objet du chapitre suivant.

CHAPITRE 3

SEGMENTATION THÉMATIQUE

Sommaire

3.1 Introduction **47**
3.2 Présentation des méthodes **47**
3.3 Paramétrage des méthodes **49**
3.3.1 Texttiling 49
3.3.2 Topictiling 52
3.3.3 Deeptiling 54
3.4 Segmentation du corpus GMMP-TV **56**
3.5 Conclusion **57**

3.1 Introduction

Dans ce chapitre, nous allons présenter les expériences réalisées sur la segmentation thématique et analyser les résultats obtenus. Nous avons choisi de comparer trois méthodes de segmentation non supervisées qui reposent sur le même principe : la détection de la rupture de similarité. Chaque méthode a été appliquée sur des corpus différents pour tenter de trouver des paramètres optimaux et mesurer leur robustesse.

3.2 Présentation des méthodes

Dans le chapitre 2, nous avons présenté les différentes méthodes de segmentation thématique. Cependant, toutes les méthodes ne sont pas adaptées pour la transcription de documents audiovisuels, notamment pour cause de manque de corpus annoté. Compte tenu de l'applicabilité et de la comparabilité des méthodes ainsi que de la complexité de la mise en place, nous avons retenu trois méthodes non supervisées pour effectuer notre expérience : Texttiling, Topictiling et Deeptiling. La raison principale de ce choix est que d'un côté, ces trois méthodes reposent sur le même principe : la détection de la rupture de cohésion, et que de l'autre côté elles se différencient par une évolution sur la modélisation du texte, passant de méthodes statistiques à des méthodes neuronales. L'expérience a pour objectif de répondre aux deux questions suivantes :

- Les méthodes de segmentation thématiques sont-elles adaptées à la transcription automatique ?

- La méthode neuronale est-elle plus performante sur la segmentation de la transcription automatique ?

Le schéma 3.1 illustre de manière générale, le processus de segmentation de ces trois méthodes. Le texte à segmenter est d'abord découpé en blocs de taille k, ensuite chaque bloc de texte est modélisé par un vecteur de valeurs numériques et enfin une courbe de similarité (figure 3.2) est tracée pour chaque paire de blocs adjacents permettant d'identifier les ruptures de similarité.

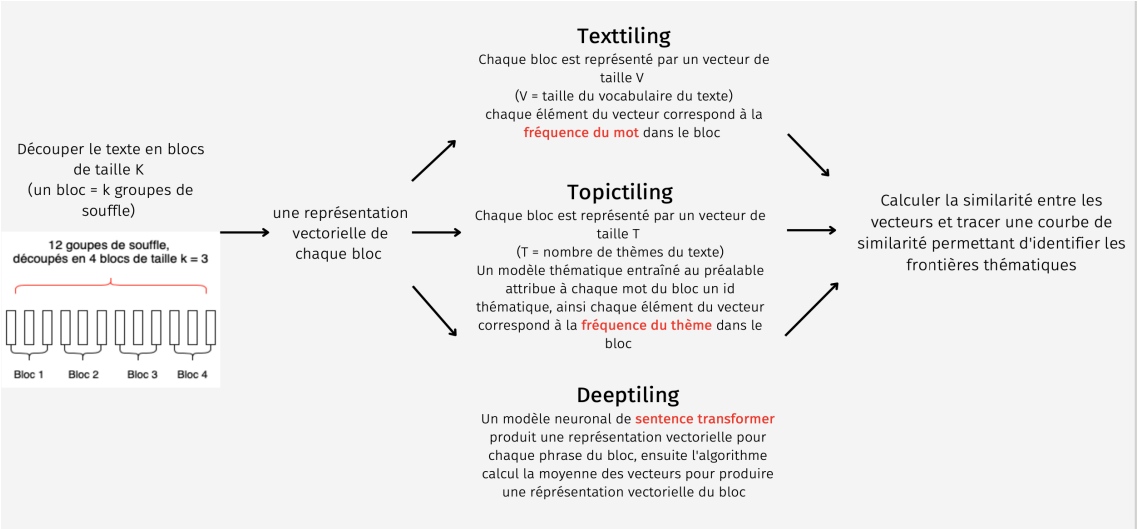


FIGURE 3.1 – Le processus de segmentation par Texttiling, Topictiling et Deeptiling

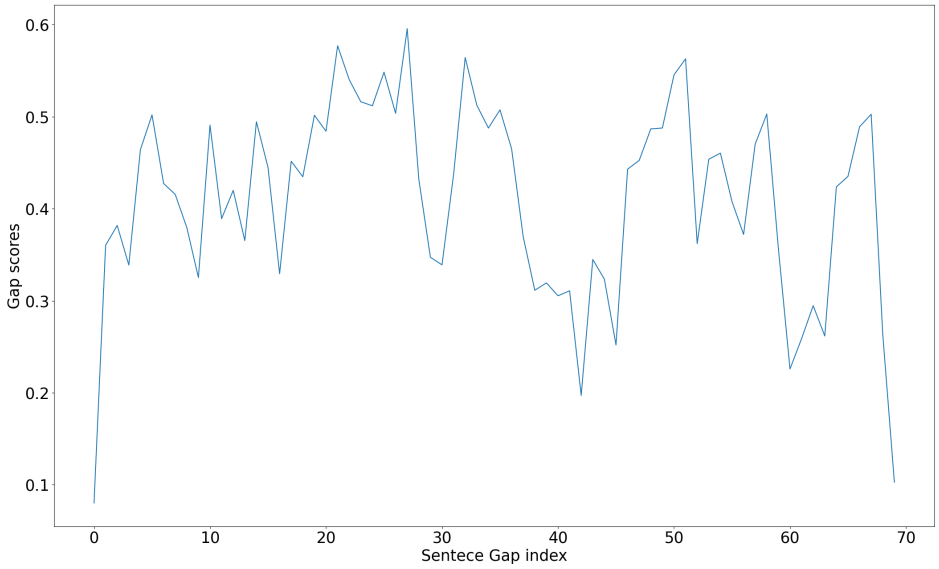


FIGURE 3.2 – Courbe de similarité entre les blocs de groupes de souffle
Gap score : la similarité cosinus entre les groupes de souffle.
Sentence Gap Index : numéro de bloc

3.3 Paramétrage des méthodes

Bien que les algorithmes choisis soient tous de type non supervisé, chacun possède certains paramètres dont l'identification d'une valeur plus ou moins idéale est nécessaire à la qualité de segmentation. Compte tenu de la similarité entre le corpus FrNewsLink et GMMP-TV, nous avons utilisé ce premier comme corpus de paramétrage et ce dernier comme corpus de test afin de comparer la performance des différentes méthodes mais aussi pour tester la robustesse des valeurs de paramètres trouvées.

3.3.1 Texttiling

La segmentation utilisant Texttiling est effectuée avec le module *texttiling* de NLTK. [1]. En entrée est fournie une transcription au format texte brut, avec un groupe de souffle par ligne. L'algorithme produit en sortie une liste de segments thématiques (Figure 3.3). Afin de faciliter l'évaluation de l'algorithme, chaque transcription segmentée est sauvegardée dans un fichier au format tabulaire dont chaque ligne contient non seulement le contenu textuel d'un groupe de souffle mais aussi les informations nécessaires à l'évaluation (Figure 3.4).

```
1  bonsoir à tout et bienvenue dans votre journal voici le titre que nous
   développeron ce soir voyager entre le martinique et le guadeloupe sous
   condition interdiction de rassemblement de plus de trente personne dont de lieu
   clore mais quel être ce lieux public claude avoir fait état ce arrêté
   préfectoral sorti hier
2  préfet et r détaille ce nouveau mesure explication dans un instant
3  le médecin cubain sur le départ mission achevée qu ' lui avoir soulevé le
   polémique alors qu ' il venir en soutien à professionnel local avoir il
   vraiment pallier le manque être ce un ouverture à d ' autre mission nous le
   verbe
4  le maison de retraite à l ' heure de le vie dix-neuf le visite être
   maintenu nous iron onze an d ' arle et voir ce qui avoir changer depuis
   le pandémie le madré port nous accueille le rédaction vous propose un long
   format ce soir et francis carole en charge de affaire sociale à le septième
   être notre invité
```

FIGURE 3.3 – Exemple d'un extrait de transcription fourni à Texttiling

Paramétrer w et k

Deux paramètres sont obligatoires pour faire fonctionner Texttiling :

- **w** : la taille (nombre de tokens) d'une pseudo-phrase (un pseudo groupe de souffle dans notre cas). Si la taille d'un groupe de souffle est supérieure à la valeur w, il est ainsi découpé en plusieurs pseudo-phrases de taille w.

- **k** : la taille d'un bloc de groupes de souffle. Cette valeur indique combien de pseudo-phrases sont à inclure dans un bloc.

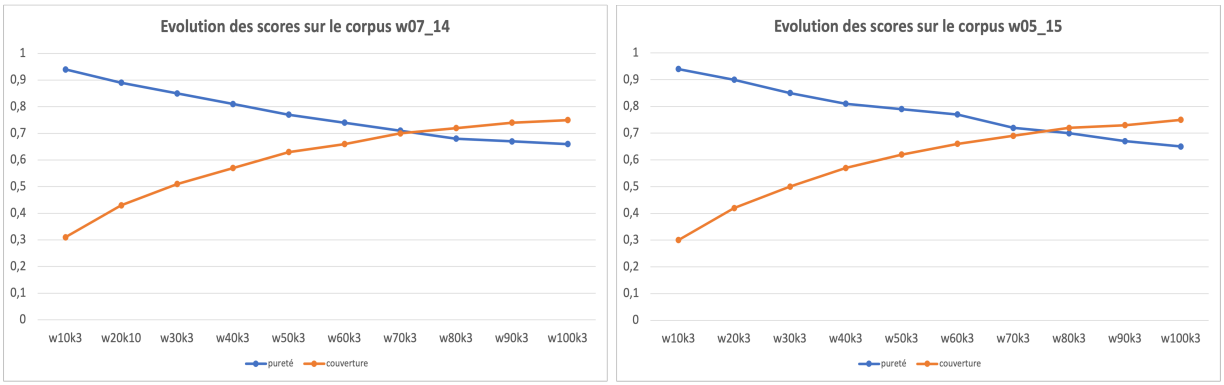
[Hearst, 1997] propose une valeur par défaut de w à 10. Afin de confirmer si cette recommandation est adaptée à notre corpus et de tester les différentes combinaisons de paramètres, nous avons utilisé le corpus W07_14 de FrNewsLink. Les premières expériences de paramétrage consistent à comparer les résultats en faisant varier la valeur de w (table 3.1). Nous avons choisi d'augmenter w de 10 à chaque fois car la taille moyenne des groupes de souffle est toujours supérieure à cette valeur. Le paramètre k est d'abord fixé à 3 car nous souhaitons que la taille de bloc soit inférieure à la taille du plus petit segment thématique.

```
1  1  8.280  25.760  bonsoir à tout et bienvenue dans votre journal voici le titre
   que nous développeron ce soir voyager entre le martinique et le guadeloupe
   sous condition interdiction de rassemblement de plus de trente personne dont de
   lieu clore mais quel être ce lieux public claude avoir fait état ce arrêté
   préfectoral sorti hier
2  0  25.760  30.600  préfet et r détaille ce nouveau mesure explication dans un
   instant
3  0  30.600  42.740  le médecin cubain sur le départ mission achevée qu ' lui
   avoir soulevé le polémique alors qu ' il venir en soutien à professionnel
   local avoir il vraiment pallier le manque être ce un ouverture à d ' autre
   mission nous le verbe
4  0  43.210  60.160  le maison de retraite à l ' heure de le vie dix-neuf le
   visite être maintenu nous iron onze an d ' arle et voir ce qui avoir
   changer depuis le pandémie le madré port nous accueille le rédaction vous
   propose un long format ce soir et francis carole en charge de affaire
   sociale à le septième être notre invité
```

FIGURE 3.4 – Exemple d'un extrait de segmentation produit par Texttiling
colonne 1 : 1 si le groupe de souffle débute un segment thématique, 0 sinon
colonne 2 : temps de début du groupe de souffle
colonne 3 : temps de fin du groupe de souffle
colonne 4 : le contenu textuel du groupe de souffle

Paramètres	Combinaisons									
w	10	20	30	40	50	60	70	80	90	100
k	3	3	3	3	3	3	3	3	3	3

TABLE 3.1 – Les paramètres de Texttiling



(a) Scores sur W07_14
bleu : pureté ; orange : couverture

(b) Scores sur W05_15
bleu : pureté ; orange : couverture

FIGURE 3.5 – Scores Texttiling avec différentes valeurs de w

La figure 3.5 (gauche) montre les scores obtenus à partir de cette première expérience, mesurés en pureté et en couverture (voir section 1.4.4). Nous pouvons constater que l'augmentation de w entraîne une baisse de pureté mais une hausse de couverture. Les scores commencent à se stabiliser quand w atteint 70. Etant donné qu"il est difficile de déterminer si la pureté ou la couverture est à privilégier, nous avons décidé de choisir un w produisant des scores plutôt équilibrés. Les paramètres w=70 et k=3 ont ainsi été retenus avec une pureté de 0.71 et une couverture de 0.70. Compte tenu de la similarité entre le corpus w05_05 et le corpus W07_14 en terme de nombre

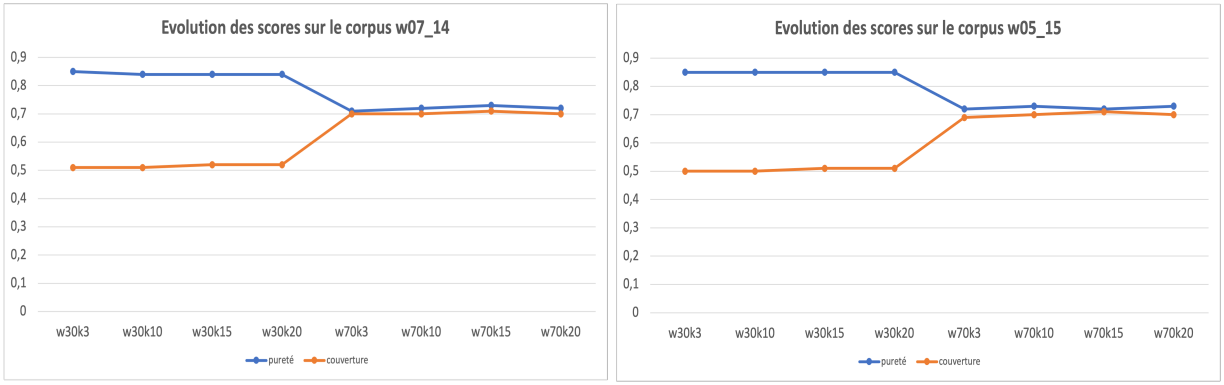
1. https://www.nltk.org/api/nltk.tokenize.texttiling.html

moyens des segments thématiques (figure 2.6), de taille moyenne des segments thématiques (figure 2.7) ainsi que de taille moyenne des groupes de souffle (figure 2.8), nous avons répété cette même expérience sur le corpus W05_15 afin de valider la pertinence de cette configuration. Les scores sont affichés dans la figure 3.5 (droite). Nous pouvons constater que les scores évoluent de manière similaire pour les deux corpus. Les paramètres w=70 et k=3 produisent une pureté de 0.72 et une couverture de 0.69 sur le corpus W05_15, assez proche de ceux obtenus sur le corpus W07_14. Au vue des expériences, il semble que les valeurs w=70 et k=3 soient les plus pertinentes.

Après avoir fixé la valeur de w, nous avons effectué une deuxième suite d'expériences pour observer si la variation du paramètre k pourrait influencer le résultat. Pour se faire, nous avons fixé deux valeurs de w : 30 et 70 et pour chaque w, nous avons testé de manière aléatoire, 4 valeurs de k. Nous introduisons tout de même w=30 dans l'expérience car nous souhaitons observer si la variation de k entraîne une variation significative des scores ou si les scores sont plutôt influencés par w. Les paramètres utilisés dans cette expérience sont résumés dans le tableau 3.2 :

Paramètres	Combinaisons							
w	30	30	30	30	70	70	70	70
k	3	10	15	20	3	10	15	20

TABLE 3.2 – Les paramètres de Texttiling



(a) Scores sur W07_14
bleu : pureté
orange : couverture

(b) Scores sur W05_15
bleu : pureté
orange : couverture

FIGURE 3.6 – Scores Texttiling avec différentes valeurs de k

Les résultats (figure 3.6) montrent que la variation de k n'impacte guère la performance de Texttiling, que ce soit en terme de pureté ou de couverture. Cependant, ceux-ci confirment à nouveau la sensibilité de l'algorithme à la variation de w car nous observons une chute significative de pureté et une augmentation importante de couverture entre w30k20 et w70k3.

Etant donné qu'une taille de fenêtre à 70 produit un résultat plutôt équilibré et que la variation de k n'apporte pas de changement significatif, nous retenons finalement w = 70 et k = 3 comme un paramètre idéal pour les expériences sur le corpus GMMP-TV.

3.3.2 Topictiling

La segmentation thématique avec Topictiling [2] est effectuée grâce à l'outil développé par [Riedl and Biemann, 2012b]. Le principe de segmentation est similaire à Texttiling mais deux étapes préliminaires sont nécessaires avant le calcul de similarité entre les vecteurs des blocs adjacents :

1. Construction d'un modèle thématique : le modèle est construit grâce à un corpus d'entraînement dont le contenu est similaire aux textes à segmenter.
2. Inférence thématique : le modèle thématique attribue un identifiant thématique à chaque mot du texte à segmenter. Par conséquent, les mots du texte sont remplacés par leurs identifiants dont la fréquence est utilisée pour convertir les blocs de textes en vecteur numérique.

Topictiling contient plusieurs paramètres, mais l'objectif du travail n'est pas de balayer toutes les combinaisons possibles, ce qui est d'ailleurs réalisé par [Riedl and Biemann, 2012a]. Ainsi, pour la plupart des paramètres, nous avons suivi les recommandations des auteurs. Notre travail de paramétrage a pour objectif de répondre aux deux questions suivantes :

1. La segmentation est-elle influencée par la variété du corpus utilisé pour entraîner un modèle thématique ?
2. Quelle est la taille de bloc idéale pour la segmentation ?

Comparer les modèles thématiques

Pour répondre à la première question, nous avons construit trois modèles thématiques avec des corpus d'entraînement issus du corpus W04_17 de FrNewsLink (tableau 3.3). Les transcriptions dans les corpus sont tous lemmatisées et les stopwords sont supprimés [3]. Le premier modèle est entraîné avec l'ensemble des programmes du W04_17 ; le deuxième modèle est entraîné avec tous les programmes de France-2, quelle que soit l'heure de diffusion ; le troisième modèle est entraîné avec sept programmes avec un programme par jour et par émission. Ces corpus d'entraînement se différencient à la fois par leurs volumes mais aussi par la variété et la quantité des thèmes contenus. Chaque modèle thématique est ensuite utilisé pour segmenter les mêmes textes.

Expérience	Corpus Train	Corpus Test
1	92 transcriptions	W05_15
2	21 transcriptions du 7h, 8h, 13h et 20h de France2	W05_15
3	7 transcriptions des émissions et des dates différentes	W05_15

TABLE 3.3 – Les différents corpus utilisés pour l'entraînement de modèles thématiques

2. https://github.com/riedlma/topictiling
3. Traitements effectués avec spaCy.

Les paramètres utilisés pour chacune des expériences sont les suivants (tableau 3.4), les valeurs de chaque paramètre ont été définies en suivant les recommandations de [Riedl and Biemann, 2012a] :

Exp	alpha	beta	m	ntopics	i	ri	mode	w
1	0.03	0.1	2000	1439	100	5	true	6
2	0.12	0.1	2000	408	100	5	true	6
3	0.56	0.1	2000	90	100	5	true	6

TABLE 3.4 – Les paramètres du Topictiling

- alpha : disparité de la distribution thème-document. Valeur recommandée : 50/T (T = nombre de thèmes du corpus d'entraînement).

- beta : disparité de la distribution thème-mot. Valeur recommandée : 0.1 - 0.01.

- m : nombre d'itérations d'estimation. Valeur recommandée : 500 - 5000.

- ntopics : nombre de thèmes du corpus d'entraînement. Valeur recommandée : 50 - 500.

- i : nombre d'itérations pour attribuer un identifiant thématique aux mots. Valeur recommandée : 100.

- ri : nombre de répétitions d'inférence. Valeur recommandée : 5.

- mode : si true, le modèle garde en mémoire l'identifiant thématique attribué à chaque mot à chaque itération. A la fin de toutes les itérations, chaque mot est annoté avec l'identifiant qui lui est plus fréquemment attribué. Valeur recommandée : true.

- w : taille du bloc de groupes de souffle. Valeur recommandée : la moitié de la taille moyenne des segments thématiques [4].

Le tableau 3.5 présente les scores moyens de l'ensemble des programmes du corpus de test. Nous constatons que le modèle thématique entraîné sur sept journaux télévisées de différentes émissions et jours semble être plus satisfaisant. Dans l'expérience 1, le corpus d'entraînement contient 1439 thèmes, ce qui dépasse l'intervalle recommandé et a produit un résultat moins satisfaisant que les autres. Bien que le nombre de thèmes pour les expériences 2 et 3 se trouvent tous dans l'intervalle recommandé, le modèle produit par l'expérience 3 atteint un score légèrement plus élevé que celui de l'expérience 2 mais avec bien moins de thèmes. On peut penser que la variété de thèmes est plus important que la quantité, car dans le corpus de l'expérience 2, plusieurs journaux d'une même journée et de la même chaîne ont été rassemblés, ce qui rend les thèmes traités moins variés que le corpus utilisé pour l'expérience 3, qui comprend seulement un journal télévisé par jour, issu de chaînes différentes.

Compte tenu de l'efficacité de calcul et la performance des modèles, nous avons décidé d'utiliser le modèle thématique construit à partir de l'expérience 3 pour les expériences ultérieures.

La taille des blocs de groupes de souffle

Le deuxième point que nous souhaitons étudier est la taille idéale du bloc (paramètre k dans Texttiling). A l'issue des premières expériences, nous avons fixé k à 3.

4. La taille moyenne des segments thématiques est de 12 pour l'ensemble de programmes du corpus W05_15.

Expériences	pureté	couverture
1	0.67	0.77
2	0.76	0.73
3	0.73	0.78

TABLE 3.5 – Scores Topictiling avec trois modèles thématiques différents

Les auteurs de Topictiling recommandent une valeur correspondant à la moitié de la taille moyenne des segments thématiques. Nous avons ainsi choisi les valeurs suivantes pour cette expérience (tableau 3.6). Nous avons choisi de tester également k = 30 car cette valeur est supérieure à la taille maximale des segments thématiques. Etant donné que le modèle thématique est entraîné sur le corpus W07_14, la segmentation est appliquée seulement sur le corpus W05_15.

Expériences	pureté	couverture
k = 3	0.70	0.71
k = 6	0.73	0.78
k = 12	0.74	0.72
k = 30	0.70	0.68

TABLE 3.6 – Scores Topictiling avec différentes valeurs de k

Nous pouvons constater que la pureté reste très proche pour différentes tailles de bloc, en revanche, la couverture oscille entre 0.68 et 0.78. Avec une taille du bloc équivalant à la moitié de la taille moyenne des segments, nous obtenons effectivement un résultat meilleur que les autres. Ces résultats montrent également qu'une taille de bloc adaptée pour Texttiling ne l'est pas forcément pour Topictiling, même si les deux méthodes partagent un même principe de segmentation. Cela pourrait s'expliquer par l'absence de paramètre w dans Topictiling : en effet, il n'est plus demandé de garantir une taille identique de groupes de souffle au sein du bloc.

Compte tenu des résultats des deux séries d'expériences ci-dessus, nous restons sur les paramètres recommandés pour la segmentation du corpus GMMP-TV, avec comme modèle thématique celui entraîné sur sept journaux télévisés issus des émissions et des dates différentes.

3.3.3 Deeptiling

La méthode Deeptiling [5] développée par [Ghinassi, 2021] est également inspirée de Texttiling, mais diffère de celui-ci sur deux aspects :

1. le vecteur représentant chaque bloc de groupes de souffle n'est pas créé à partir du calcul de fréquence des mots dans le corpus mais extrait d'un encodeur neuronal de phrase pré-entraîné. L'encodeur utilisé dans notre expérience est *paraphrase-xlm-r-multilingual-v1* [6], un modèle a été choisi car il est capable de

5. https://github.com/Ighina/DeepTiling
6. https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

produire des plongements pour des textes multilingues grâce au processus de distillation de connaissances (Knowledge distillation). Chaque groupe de souffle du bloc est modélisé par un vecteur de taille 768. Ensuite on calcule la moyenne des vecteurs des phrases au sein du bloc pour produire un seul vecteur représentant celui-ci.

2. Etant donné que la dimension de chaque vecteur de phrase est identique, il n'est pas nécessaire de définir une taille de pseudo-phrase comme dans Texttiling.

L'algorithme Deeptiling a également comme paramètre la taille de bloc. Compte tenu de la comparabilité avec les deux autres algorithmes, nous avons conduit l'expérience avec les mêmes valeurs de k testées pour Topictiling, ce qui donne les scores suivants (tableau 3.7) :

Expérience	pureté	couverture
k = 3	0.80	0.74
k = 6	0.79	0.83
k = 12	0.73	0.92
k = 30	0.66	0.90

TABLE 3.7 – Scores Deeptiling avec différentes valeurs de k

Nous pouvons constater qu'une taille de bloc trop petite ou trop grande fait baisser respectivement la couverture et la pureté et que les scores sont relativement stables pour k = 6 et k = 12. Bien que k = 6 apporte une pureté plus élevée, sa couverture reste bien inférieure à la couverture produite par k = 12. Nous souhaitons privilégier une sous-segmentation car il est plus simple d'ajouter des frontières thématiques que de rassembler les segments trop morcelés. Nous retenons ainsi la taille moyenne des segments comme taille de bloc.

Le tableau 3.8 résume la pureté et la couverture obtenus sur le corpus W05_15 par chacun des trois algorithmes avec les paramètres retenus :

Algorithme	pureté	couverture
Texttiling (w=70, k=3)	0.72	0.69
Topictiling (k=6)	0.73	0.78
Deeptiling (k=12)	0.73	0.92

TABLE 3.8 – Scores obtenus par différents algorithmes de segmentation sur le corpus W05_15

La performance des trois algorithmes est plutôt similaire en terme de pureté, en revanche, une nette amélioration est observée chez Deeptiling en terme de couverture. Cela signifie que Deeptiling éprouve des difficultés quant à l'identification du nombre adéquat de segments, c'est-à-dire qu'il a tendance à regrouper des segments. Cependant, les frontières thématiques sont pour la plupart correctement placées.

3.4 Segmentation du corpus GMMP-TV

Dans la section précédente, nous avons présenté le travail de paramétrage des algorithmes. Les paramètres retenus pour les trois algorithmes sont les suivants :

- Texttiling : w = 70, k = 3

- Topictiling : k = la moitié de la taille moyenne des segments thématiques des documents. Le modèle thématique utilisé est celui entraîné sur un corpus de sept journaux télévisés d'émissions et de dates différentes.

- Deeptiling : k = la taille moyenne des segments thématiques des documents.

Chacune des méthodes a été appliquée sur le corpus GMMP-TV et les scores moyens sont résumés dans le tableau 3.9 :

Expérience	pureté	couverture
Texttiling (w=70, k=3)	0.70	0.69
Topictiling (k=7)	0.77	0.68
Deeptiling (k=14)	0.73	0.96

TABLE 3.9 – Scores obtenus par différents algorithmes de segmentation sur le corpus GMMP-TV

Nous pouvons constater que la pureté est plutôt stable quelle que soit la méthode, mais la couverture est nettement plus élevée avec Deeptiling. Cela revèle une réelle évolution de performance puisque la couverture mesure le taux de couverture de segment de référence par rapport au segment obtenu automatiquement. Globalement, les scores produits ressemblent à ceux obtenus sur la segmentation du corpus W05_15, ce qui confirme notre choix de paramètres. Il est à noter que Topictiling ne se comporte pas de la même manière sur le corpus W05_15 que sur GMMP-TV. Cela montre que l'introduction du modèle thématique a un réel impact sur la segmentation. En effet, la segmentation pourrait être plus ou moins satisfaisante en fonction du niveau de similarité du contenu entre le corpus utilisé pour l'entraînement du modèle thématique et celui à segmenter. La capacité de généralisation du Topictiling est ainsi limitée car les journaux télévisés suivent l'actualité qui est en constante évolution, rendant ainsi difficile de construire un modèle thématique assez général pour être performant sur des corpus de périodes différents.

Si l'on se concentre sur les scores produits sur GMMP-TV, on peut observer que la pureté est supérieure à la couverture avec Texttiling et Topictiling, mais Deeptiling inverse totalement le rapport entre les deux métriques. Entre Texttiling et Topictiling, chacun retrouve, pour chaque segment de référence, environ 70% de textes correspondant, cela signifie que les frontières thématiques identifiées par ces deux méthodes ne s'approchent pas beaucoup de celles attendues. Deeptiling, par rapport aux deux autres méthodes, est plus performant pour poser les frontières aux emplacements corrects ou quasi-corrects, bien que la pureté relativement faible révèle la présence d'une sous-segmentation.

La figure 3.7 illustre les scores obtenus pour chaque programme du corpus GMMP-TV par les trois algorithmes. Il convient de noter que parmi les trois algorithmes, Texttiling et Topictiling présentent des performances semblables, que ce soit en terme de pureté et de couverture, généralement comprises entre 0.6 et 0.8. Deeptiling semble être le plus performant car, bien qu'ayant une pureté comparable

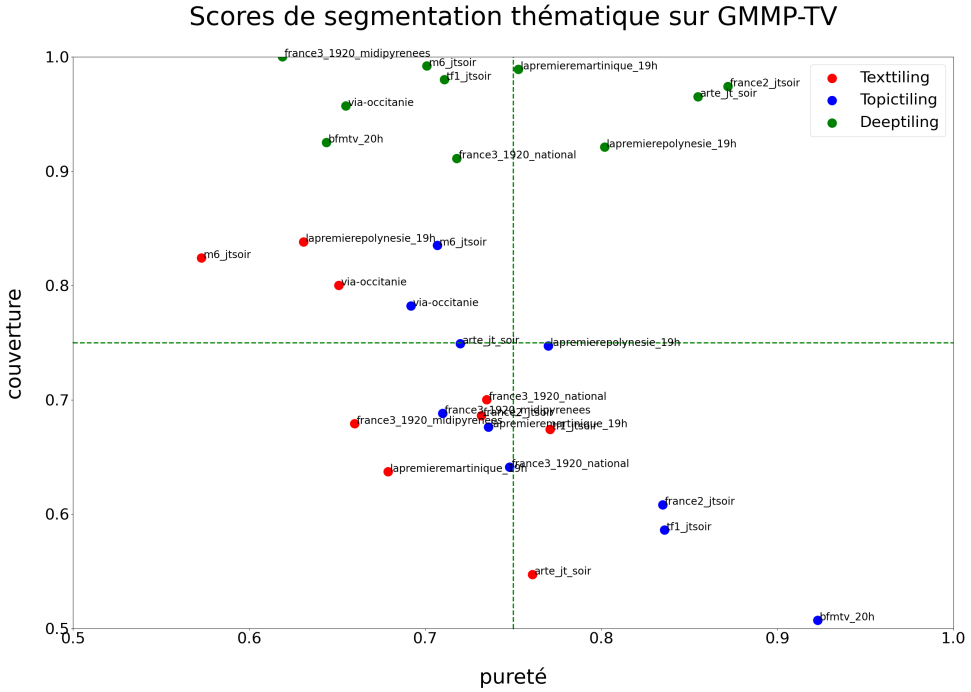


FIGURE 3.7 – Les scores de segmentation thématique sur GMMP-TV

aux deux autres algorithmes, présente une bien meilleure couverture. De plus, seul Deeptiling arrive à se positionner dans la partie haute du diagramme, synonyme de haute pureté et couverture.

3.5 Conclusion

Dans ce chapitre, nous avons présenté les expériences de segmentation thématique réalisées avec trois méthodes différentes, qui sont initialement utilisées sur des textes écrits. Les résultats de nos expériences montrent que ces méthodes sont relativement performantes sur les transcriptions de journaux télévisés, bien que celles-ci contiennent des erreurs de transcriptions. Les expériences confirment également la capacité supérieure de la méthode neuronale (Deeptiling) : utiliser les plongements de phrases à partir du modèle de langues pré-entraîné a permis de produire une segmentation bien plus proche des résultats attendus. En terme de facilité d'utilisation, il y a moins de paramètres à gérer pour Deeptiling, en particulier par rapport à Topictiling qui nécessite un entraînement au préalable d'un modèle thématique, sur un corpus similaire au niveau de contenu.

Bien évidemment, aucune de ces méthodes n'est sans faille. Quelle que soit la méthode utilisée, il est préférable de connaître en avance au moins le nombre de segments afin de mieux définir la taille de bloc, un paramètre commun des trois méthodes. Or, il n'est pas toujours possible d'obtenir cette information, ce qui implique de multiplier les essais afin d'atteindre un résultat satisfaisant.

CHAPITRE 4

PERSPECTIVES

Sommaire

4.1 Introduction **59**
4.2 Améliorer la segmentation thématique **59**
4.3 Identification thématique **61**
4.4 Conclusion **62**

4.1 Introduction

Dans ce chapitre, nous allons tenter d'ouvrir des perspectives sur la tâche de segmentation thématique. Nous allons tout d'abord tenter de proposer des pistes de réflexion qui pourraient permettre d'améliorer la segmentation, puis nous présenterons une tâche possible suite à la segmentation thématique qu'est l'identification thématique ainsi que certains points devant encore être explorés.

4.2 Améliorer la segmentation thématique

Dans le chapitre précédent, nous avons appliqué trois méthodes différentes de segmentation thématique aux transcriptions automatiques des journaux télévisés. La performance de ces trois méthodes demeure stable sur des corpus de contenu différents, avec notamment la méthode neuronale qui obtient une couverture assez satisfaisante. Cependant, la capacité de généralisation des méthodes reste à étudier car les journaux télévisés ne sont pas représentatifs de toutes les données audiovisuelles. Les journaux télévisés sont souvent plus structurés comparé à d'autres émissions, par exemple les libres antennes pour lesquelles, même pour un annotateur humain, il peut s'avérer être compliqué d'identifier des changements thématiques. En effet, dans ce type d'émissions, les échanges sont plus vifs, entraînant ainsi des interruptions de paroles. De plus, les paroles spontanées y sont plus nombreuses, les locuteurs peuvent entre deux propos sur un sujet donné, effectuer une aparté sur un sujet qui n'a aucun rapport avec ce dernier.

L'extrait ci-dessous (figure 4.1), issu de l'émission « After foot », est révélateur des problèmes mentionnés auparavant. On peut remarquer qu'à plusieurs endroits, les transcriptions n'ont pas de sens (soulignés en rouge) : lorsque l'on écoute l'émission, on remarque que cela se produit au moment où plusieurs personnes parlent simultanément, ou bien lorsque le locuteur n'articule pas suffisamment. Ces erreurs lors

de la transcription deviennent réellement problématiques quand certains termes, habituellement annonciateurs d'une transition entre plusieurs sujets sont manquants ou non correctement retranscrits. Par exemple, la phrase soulignée en jaune, « Alors chers afteriens, vous qui êtes branchés sur RMC sport... » est révélatrice en temps normal d'un changement de sujet car, lors des émissions radio, ce type de phrase où l'on rappelle la station sur laquelle on se trouve peut signifier une reprise d'antenne ou un changement de sujet. Or ici, le fait que la transcription ait été erronée implique qu'il devient plus difficile de déterminer les frontières entre les différents sujets.

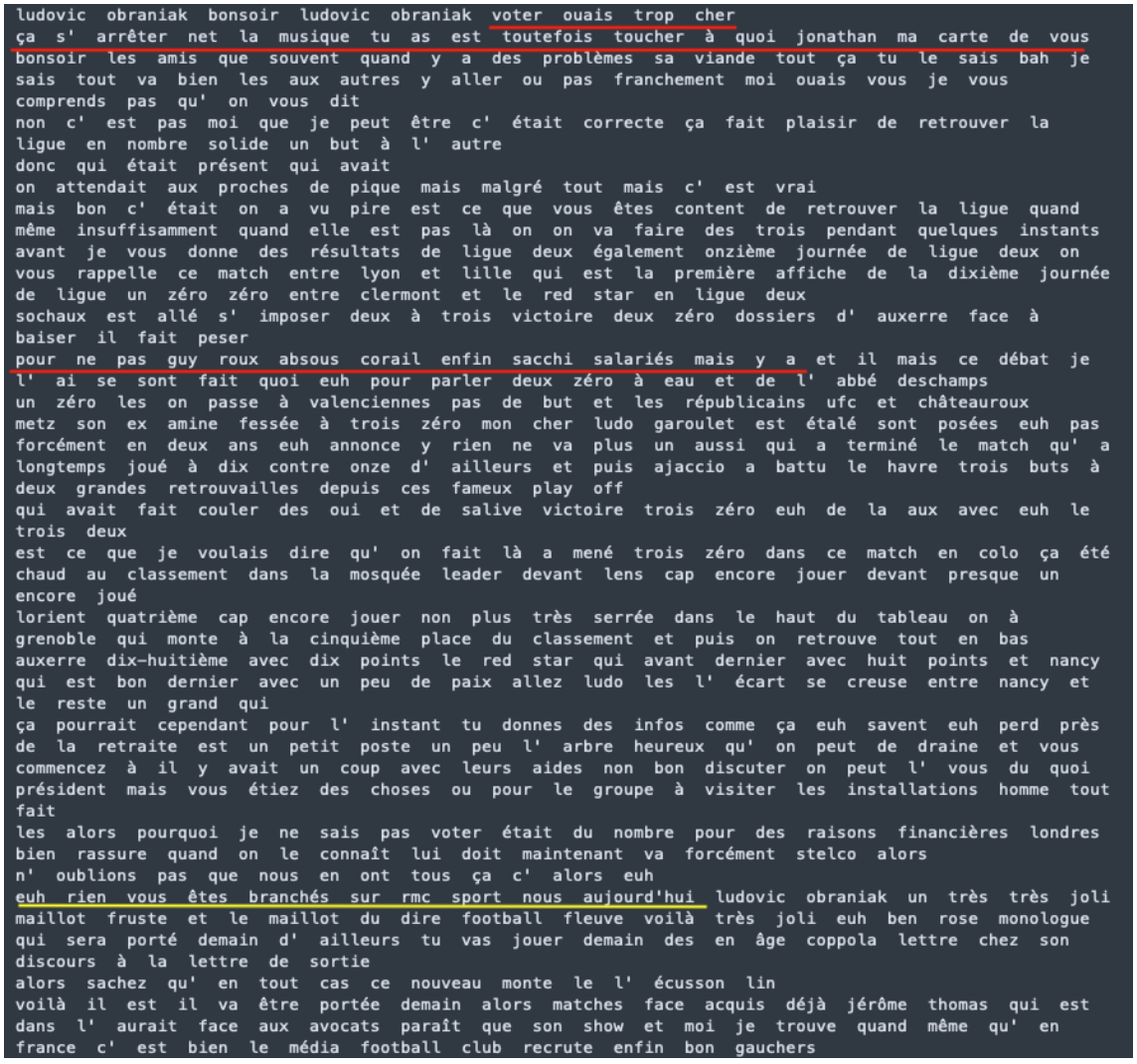

ludovic obraniak bonsoir ludovic obraniak voter ouais trop cher
ça s' arrêter net la musique tu as est toutefois toucher à quoi jonathan ma carte de vous
bonsoir les amis que souvent quand y a des problèmes sa viande tout ça tu le sais bah je
sais tout va bien les aux autres y aller ou pas franchement moi ouais vous je vous
comprends pas qu' on vous dit
non c' est pas moi que je peut être c' était correcte ça fait plaisir de retrouver la
ligue en nombre solide un but à l' autre
donc qui était présent qui avait
on attendait aux proches de pique mais malgré tout mais c' est vrai
mais bon c' était on a vu pire est ce que vous êtes content de retrouver la ligue quand
même insuffisamment quand elle est pas là on on va faire des trois pendant quelques instants
avant je vous donne des résultats de ligue deux également onzième journée de ligue deux on
vous rappelle ce match entre lyon et lille qui est la première affiche de la dixième journée
de ligue un zéro zéro entre clermont et le red star en ligue deux
sochaux est allé s' imposer deux à trois victoire deux zéro dossiers d' auxerre face à
baiser il fait peser
pour ne pas guy roux absous corail enfin sacchi salariés mais y a et il mais ce débat je
l' ai se sont fait quoi euh pour parler deux zéro à eau et de l' abbé deschamps
un zéro les on passe à valenciennes pas de but et les républicains ufc et châteauroux
metz son ex amine fessée à trois zéro mon cher ludo garoulet est étalé sont posées euh pas
forcément en deux ans euh annonce y rien ne va plus un aussi qui a terminé le match qu' a
longtemps joué à dix contre onze d' ailleurs et puis ajaccio a battu le havre trois buts à
deux grandes retrouvailles depuis ces fameux play off
qui avait fait couler des oui et de salive victoire trois zéro euh de la aux avec euh le
trois deux
est ce que je voulais dire qu' on fait là a mené trois zéro dans ce match en colo ça été
chaud au classement dans la mosquée leader devant lens cap encore jouer devant presque un
encore joué
lorient quatrième cap encore jouer non plus très serrée dans le haut du tableau on à
grenoble qui monte à la cinquième place du classement et puis on retrouve tout en bas
auxerre dix-huitième avec dix points le red star qui avant dernier avec huit points et nancy
qui est bon dernier avec un peu de paix allez ludo les l' écart se creuse entre nancy et
le reste un grand qui
ça pourrait cependant pour l' instant tu donnes des infos comme ça euh savent euh perd près
de la retraite est un petit poste un peu l' arbre heureux qu' on peut de draine et vous
commencez à il y avait un coup avec leurs aides non bon discuter on peut l' vous du quoi
président mais vous étiez des choses ou pour le groupe à visiter les installations homme tout
fait
les alors pourquoi je ne sais pas voter était du nombre pour des raisons financières londres
bien rassure quand on le connaît lui doit maintenant va forcément stelco alors
n' oublions pas que nous en ont tous ça c' alors euh
euh rien vous êtes branchés sur rmc sport nous aujourd'hui ludovic obraniak un très très joli
maillot fruste et le maillot du dire football fleuve voilà très joli euh ben rose monologue
qui sera porté demain d' ailleurs tu vas jouer demain des en âge coppola lettre chez son
discours à la lettre de sortie
alors sachez qu' en tout cas ce nouveau monte le l' écusson lin
voilà il est il va être portée demain alors matches face acquis déjà jérôme thomas qui est
dans l' aurait face aux avocats paraît que son show et moi je trouve quand même qu' en
france c' est bien le média football club recrute enfin bon gauchers

FIGURE 4.1 – L'extrait du programme After foot

Par conséquence, si l'on souhaite entraîner un modèle surpervisé basé sur la transcription automatique pour segmenter ce type d'émission, il sera plus compliqué de fournir un corpus d'apprentissage de qualité car cela nécessite non seulement une annotation mais aussi une correction manuelle de la transcription. Compte tenu de la performance de la méthode neuronale sur la segmentation des journaux télévisés, deux pistes d'amélioration pourraient être envisagées :

1. Tester les modèles transformers pour la transcription automatique et évaluer l'évolution du taux d'erreurs mais également la performance sur des mots hors vocabulaire (OOV). Nous pouvons par exemple utiliser les corpus audio à notre

disposition afin d'affiner le modèle wav2vec 2.0 [Baevski et al., 2020], un modèle utilisant l'apprentissage auto-supervisé.

2. Entraîner des modèles thématiques à l'aide de corpus diachroniques : on peut supposer que lors d'une même journée les sujets d'actualité abordés, et ce quelque soit le canal utilisé (presse écrite, radio, télévision, ...) seront à de rares exceptions près sensiblement les mêmes. Nous pouvons ainsi, pour une période donnée, entraîner un modèle thématique sur la presse écrite, puis utiliser ce modèle afin de segmenter les émissions radiophoniques ou audiovisuelles de la même période.

4.3 Identification thématique

Nous avons mentionné précédemment que la segmentation thématique est souvent considérée comme un travail préliminaire aux autres tâches de traitement automatique des langues, comme par exemple l'indexation des documents. Une segmentation efficace est certes utile, mais le travail serait plus complet si l'on parvenait à connaître le thème de chaque segment obtenu. Pour ce faire, nous avons tenté d'extraire les thèmes de chaque segment automatiquement obtenu en y appliquant un modèle thématique.

Le tableau 4.1 résume le thème de chaque segment thématique de référence pour le programme *arte_jt_soir*. Nous avons choisi de tester celui-ci car le résultat de la segmentation automatique se rapproche le plus de la référence.

Segments	Thèmes
1	Annonce des titres
2	Election présidentielle américaine
3	5G en France
4	Tensions en Azerbaïdjan
5	Manifestations en Catalogne
6	Impacts économiques du Covid-19 sur l'Egypte
7	Jessica Auer : photographier pour faire appel à la protection de l'environnement
8	Fin du journal

TABLE 4.1 – Le thème de chaque segment thématique du programme *arte_jt_soir* (référence)

La figure 4.2 illustre les mots les plus fréquents appartenant aux thèmes obtenus sur le programme *arte_jt_soir*. Dans la segmentation de référence, les cinq premiers segments correspondent exactement à la référence, en revanche, le segment 6 contient un seul groupe de souffle et le segment 7 regroupe les trois derniers segments de référence. Le modèle thématique utilisé pour l'inférence des thèmes est celui entraîné pour l'expérience Topictiling (voir section 3.3.2). Etant donné que le modèle a été entraîné sur 90 thèmes, au moment d'inférence, il produit également

seg1	seg2	seg3	seg4	seg5	seg6	seg7
56 débat	22 dizaine	20 radio	26 mort	21 demi	77 jeune	39 augmentation
81 programme	44 passer	37 onde	50 bilan	43 an	86 père	68 oui
85 bienvenue	54 opposer	62 devoir	68 majorité	60 cour	88 cas	70 prix
86 arte	66 combat	72 clair	71 attaque	69 président	89 alimentation	76 bien
88 côte	71 envie	77 enchère	79 livre	74 annonce	89 majorité	79 aller
89 éric	75 presque	78 lancer	83 combat	83 connaître	89 vide	81 euh
89 suisse	79 débat	79 france	83 jour	84 hier	90 changer	
90 année	80 grand	79 gramme	83 poursuivre	84 manifestation		
	81 heure	80 huile				
		82 lampe				

FIGURE 4.2 – Le thème de chaque segment thématique du programme *arte_jt_soir* (hypothèse)

90 thèmes d'hypothèses pour chaque segment. Ainsi, « 90 année » signifie que que le mot « année » est apparu dans chaque thème inféré. Pour chaque segment, les mots en rouge reflètent plus ou moins le thème de celui-ci : par exemple, pour le segment 2, on peut relever que l'un des mots en rouge est « débat », ce qui est cohérent avec le thème des élections américaines. Il en va de même avec le thème de la 5G en France dont l'un des mots récurrent est « onde ». Cependant, on peut souligner que certains mots clé du thème ne se trouvent pas dans la liste de mots. Par exemple pour le thème « élection présidentielle américaine », on peut s'attendre à voir des mots comme « Joe Biden », « Donald Trump », « Amérique », « Etats-Unis » etc. qui ne sont pourtant pas dans la liste des mots fréquents. Ce phénomène peut s'expliquer par le fait que le modèle thématique ne reconnaît pas les mots n'apparaissant pas dans le corpus d'entraînement, ce qui est notamment le cas pour « Joe Biden », « Donald Trump » ou encore « covid ». De la même manière, tous les mots récurrents (ici en bleu) ne sont pas tous porteurs d'information thématique. Par exemple, le mot « suisse » a une fréquence très élevée, cependant, il apparaît seulement dans la publicité avant le début du journal. Pour les travaux futurs, peut-être faudrait-il penser à explorer d'autres méthodes plus récentes de modélisation thématique avec les documents automatiquement segmentés. Si les segments sont correctement obtenus, on peut s'attendre à obtenir le même nombre de thèmes avec des mots fréquents reflétant mieux le sujet abordé. On pourrait également observer si d'autres méthodes pourraient permettre d'éviter les problèmes rencontrés ci-dessus et si le segment pourrait éventuellement encore être divisé.

4.4 Conclusion

La segmentation thématique est une tâche complexe et les méthodes varient en fonction du contenu à segmenter. Dans ce chapitre, nous avons proposé quelques travaux à réaliser dans le futur afin d'améliorer les résultats obtenus à partir de nos expériences. Nous avons également introduit une tâche en aval qui pourrait compléter le travail de segmentation thématique.

Conclusion générale

La segmentation thématique est l'une des tâches étudiées dans le domaine du traitement automatique des langues (TAL). Segmenter les documents en thèmes permet de faciliter la recherche d'informations au milieu de données volumineuses et peut aussi épauler les études dans les autres domaines en leur fournissant un outil de travail.

De nombreuses méthodes ont été développées pour la segmentation thématique automatique. Cependant, certaines sont contraintes par le corpus d'entraînement à disposition, d'autres sont applicables sur un certain type de données. De plus, l'évaluation d'un tel modèle n'est pas évident car celle-ci subit une subjectivité à double sens : d'un côté, l'emplacement de la frontière thématique peut varier d'une personne à une autre car le thème est par nature un concept difficile à définir. De l'autre côté, les métriques d'évaluation conçus nécessitent pour la plupart d'introduire une valeur subjective afin de comparer la référence à l'hypothèse. Cela nous permet de prendre du recul vis-à-vis des résultats numériques obtenus et de prendre en compte des spécificités de données quant au choix des outils.

Dans ce travail, nous nous intéressons à la segmentation thématique de transcriptions de données audiovisuelles. La transcription, à la différence des textes classiques, contient des caractéristiques qui pourraient impacter la qualité de la segmentation. Sur la forme, un texte transcrit n'a pas de ponctuation et n'est pas structuré en phrases ni en paragraphes. Sur le fond, les erreurs de transcriptions pourraient être plus ou moins importantes en fonction de la qualité du système de reconnaissance de parole utilisé. De plus, certaines émissions audiovisuelles contiennent davantage de paroles spontanées, dont l'identification du thème pourrait être difficile même pour un humain.

L'objectif de ce travail est d'analyser si les méthodes de segmentation performantes sur du document écrit peuvent aussi être efficaces sur des données orales. Pour ce faire, nous avons choisi trois méthodes qui représentent en quelque sorte l'évolution du TAL, en passant des méthodes statistiques (Texttiling, Topictiling) à la méthode neuronale (Deeptiling). Nous proposons par ailleurs d'évaluer la segmentation thématique en utilisant un métrique initialement appliqué pour la segmentation en locuteur. Nous estimons que ce métrique est relativement objectif car il ne nécessite pas de définir une tolérance aux frontières de segmentation. Les résultats des expériences montrent que la performance des deux méthodes statistiques, à savoir Texttiling et Topictiling, sont semblables, avec des puretés respectives de 0.7 et 0.77 ainsi que des couvertures respectives de 0.69 et 0.68. La méthode neuronale produit une couverture beaucoup plus élevée, atteignant 0.96, ce qui signifie que toutes les frontières thématiques retrouvées sont correctes ou quasi-correctes. En revanche, Deeptiling est handicapé par une pureté plus faible (0.73), ce qui pourrait se révéler problématique si nous souhaitons obtenir une segmentation plus fine.

Notre travail présente également quelques limites : tout d'abord, les méthodes choisies nécessitent de configurer plusieurs paramètres et nous n'avons pas pu cou-

vrir plus de combinaisons dans nos expériences. De plus, chacune des méthodes nécessite au moins de définir au préalable la taille du bloc de comparaison, or il est difficile de trouver une valeur idéale pour celle-ci car la taille d'un segment thématique peut varier d'un document à l'autre. De plus, nous avons remarqué la présence d'erreurs de transcription dans notre corpus. Cependant, il est difficile de conclure si la performance des méthodes a été impactée par ces erreurs puisque nous ne pouvons pas comparer les résultats obtenus avec ceux obtenus sur un corpus proprement transcrit. En revanche, nous avons observé que les erreurs pourraient avoir un impact sur l'identification thématique des segments automatiquement obtenus, du moins si l'on utilise un modèle thématique statistique. Enfin, les expériences se sont limitées aux transcriptions des journaux télévisées car nous ne pouvons pas évaluer la segmentation des émissions radiophoniques, faute de corpus de référence. Toutes ces limites nous montrent que le potentiel de la segmentation thématique n'est pas encore pleinement exploité. A l'avenir, peut-être pourrions nous non seulement essayer de compléter les expériences à partir de ces limites identifiées, mais aussi essayer de varier les méthodes pour chaque étape dans la chaîne de travail : par exemple, nous pourrions identifier d'autres méthodes de reconnaissance automatique de paroles afin de comparer si nous obtenons de meilleures transcriptions ; entraîner un modèle thématique avec un corpus de journaux écrits produits à la même période que les journaux télévisés à segmenter.

BIBLIOGRAPHIE

[Allan et al., 1998] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic Detection and Tracking Pilot Study Final Report. – Cité page 19.

[Amaral and Trancoso, 2003] Amaral, R. and Trancoso, I. (2003). Topic indexing of tv broadcast news programs. In Mamede, N. J., Trancoso, I., Baptista, J., and das Graças Volpe Nunes, M., editors, *Computational Processing of the Portuguese Language*, pages 219–226, Berlin, Heidelberg. Springer Berlin Heidelberg. – Cité page 19.

[Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc. – Cité page 61.

[Barzilay and Elhadad, 1997] Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization*. – Cité page 19.

[Beeferman et al., 1999] Beeferman, D., Berger, A., and Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34:177–210. – Cité page 26.

[Blei et al., 2001] Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press. – Cité page 22.

[Bouchekif, 2016] Bouchekif, A. (2016). *Structuration automatique de documents audio*. Theses, Université du Maine. – Cité page 19.

[Bouchekif et al., 2014] Bouchekif, A., Damnati, G., and Charlet, D. (2014). Speech cohesion for topic segmentation of spoken contents. pages 1890–1894. – Cité page 25.

[Bouchekif et al., 2015] Bouchekif, A., Damnati, G., Estève, Y., Charlet, D., and Camelin, N. (2015). Diachronic Semantic Cohesion for Topic Segmentation of TV Broadcast News. In *Interspeech 2015*, Dresden, Germany. – Cité page 24.

[Brémond and Pavel, 1988] Brémond, C. and Pavel, T. (1988). La fin d'un anathème. *Communications*, 47(1):209–220. Included in a thematic issue : Variations sur le thème. Pour une thématique. – Cité page 18.

[Camelin et al., 2018] Camelin, N., Damnati, G., Bouchekif, A., Landeau, A., Charlet, D., and Estève, Y. (2018). FrNewsLink : a corpus linking TV Broadcast News Segments and Press Articles. In *LREC 2018*, Miyazaki, Japan. – Cité pages 34 et 39.

[Choi, 2000] Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. – Cité page 21.

[Collot, 1988] Collot, M. (1988). Le thème selon la critique thématique. *Communications*, 47(1):79–91. Included in a thematic issue : Variations sur le thème. Pour une thématique. – Cité page 18.

[Coulomb-Gully, 2012] Coulomb-Gully, M. (2012). Genre et médias : vers un état des lieux. *Sciences de la société*, 83:3–13. – Cité page 37.

[Dias et al., 2007] Dias, G., Alves, E., and Lopes, J. G. P. (2007). Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1334–1339. AAAI Press. – Cité page 19.

[Dumont and Quénot, 2012] Dumont, E. and Quénot, G. (2012). Automatic Story Segmentation for TV News Video Using Multiple Modalities. *International Journal of Digital Multimedia Broadcasting*, 2012:Article ID 732514, 11p. – Cité page 25.

[Galley et al., 2003] Galley, M., McKeown, K. R., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics. – Cité pages 20, 21, 24 et 25.

[Georgescul et al., 2006] Georgescul, M., Clark, A., and Armstrong, S. (2006). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 144–151, Sydney, Australia. Association for Computational Linguistics. – Cité page 24.

[Ghinassi, 2021] Ghinassi, I. (2021). Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. unsupervised text segmentation via deep sentence encoders. – Cité pages 24 et 54.

[Glavaš and Somasundaran, 2020] Glavaš, G. and Somasundaran, S. (2020). Two-level transformer and auxiliary coherence modeling for improved text segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7797–7804. – Cité page 23.

[Guinaudeau, 2011] Guinaudeau, C. (2011). *Structuration automatique de flux télévisuels*. Theses, INSA de Rennes. – Cité page 19.

[Guinaudeau et al., 2012] Guinaudeau, C., Gravier, G., and Sébillot, P. (2012). Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech and Language*, 26(2):90–104. – Cité pages 24 et 26.

[Halliday and Hasan, 1976] Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London. – Cité page 21.

[Hauptmann and Witbrock, 1998] Hauptmann, A. and Witbrock, M. (1998). Story segmentation and detection of commercials in broadcast news video. In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98-*, pages 168–179. – Cité page 25.

[Hearst, 1997] Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64. – Cité pages 19, 21 et 49.

[Hmayda et al., 2020] Hmayda, M., Ejbali, R., and Zaied, M. (2020). Classification program and story boundaries segmentation in tv news broadcast videos via deep convolutional neural network. *Journal of Computer Science*, 16(5):601–619. – Cité page 25.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. – Cité page 23.

[Koshorek et al., 2018] Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. (2018). Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics. – Cité pages 22 et 23.

[Lamprier et al., 2007] Lamprier, S., Amghar, T., Levrat, B., and Saubion, F. (2007). On evaluation methodologies for text segmentation algorithms. volume 2, pages 19–26. – Cité page 28.

[Meignier and Merlin, 2010] Meignier, S. and Merlin, T. (2010). LIUM SPKDIARIZATION: AN OPEN SOURCE TOOLKIT FOR DIARIZATION. In *CMU SPUD Workshop*, Dallas, United States. – Cité page 34.

[Misra et al., 2009] Misra, H., Yvon, F., Jose, J. M., and Cappe, O. (2009). Text segmentation via topic modeling: An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 1553–1556, New York, NY, USA. Association for Computing Machinery. – Cité page 22.

[Morris and Hirst, 1991] Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48. – Cité page 21.

[Nie et al., 2013] Nie, X., Feng, W., Wan, L., and Xie, L. (2013). Measuring semantic similarity by contextualword connections in chinese news story segmentation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8312–8316. – Cité page 24.

[Pevzner and Hearst, 2002] Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36. – Cité page 27.

[Rastier, 1995] Rastier, F. (1995). La sémantique des thèmes-ou le voyage sentimental. *L'analyse thématique des données textuelles. L'exemple des sentiments, Paris: Didier*, pages 223–249. – Cité pages 18 et 19.

[Reynar, 1994] Reynar, J. C. (1994). An automatic method of finding topic boundaries. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 331–333, Las Cruces, New Mexico, USA. Association for Computational Linguistics. – Cité page 21.

[Riedl and Biemann, 2012a] Riedl, M. and Biemann, C. (2012a). Sweeping through the topic space: Bad luck? roll again! In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 19–27, Avignon, France. Association for Computational Linguistics. – Cité pages 52 et 53.

[Riedl and Biemann, 2012b] Riedl, M. and Biemann, C. (2012b). TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics. – Cité pages 22 et 52.

[Sheikh et al., 2017] Sheikh, I., Fohr, D., and Illina, I. (2017). Topic segmentation in ASR transcripts using bidirectional rnns for change detection. In *ASRU 2017 - IEEE Automatic Speech Recognition and Understanding Workshop*, proceedings of IEEE ASRU 2017, Okinawa, Japan. – Cité page 24.

[Sparck Jones, 1972] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. – Cité page 22.

[Stockinger, 2003] Stockinger, P. (2003). *Le document audiovisuel : procédures de description et exploitation*. Hermes Science Publications. – Cité page 18.

[Sun et al., 2008] Sun, Q., Li, R., Luo, D., and Wu, X. (2008). Text segmentation with lda-based fisher kernel. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, page 269–272, USA. Association for Computational Linguistics. – Cité page 22.

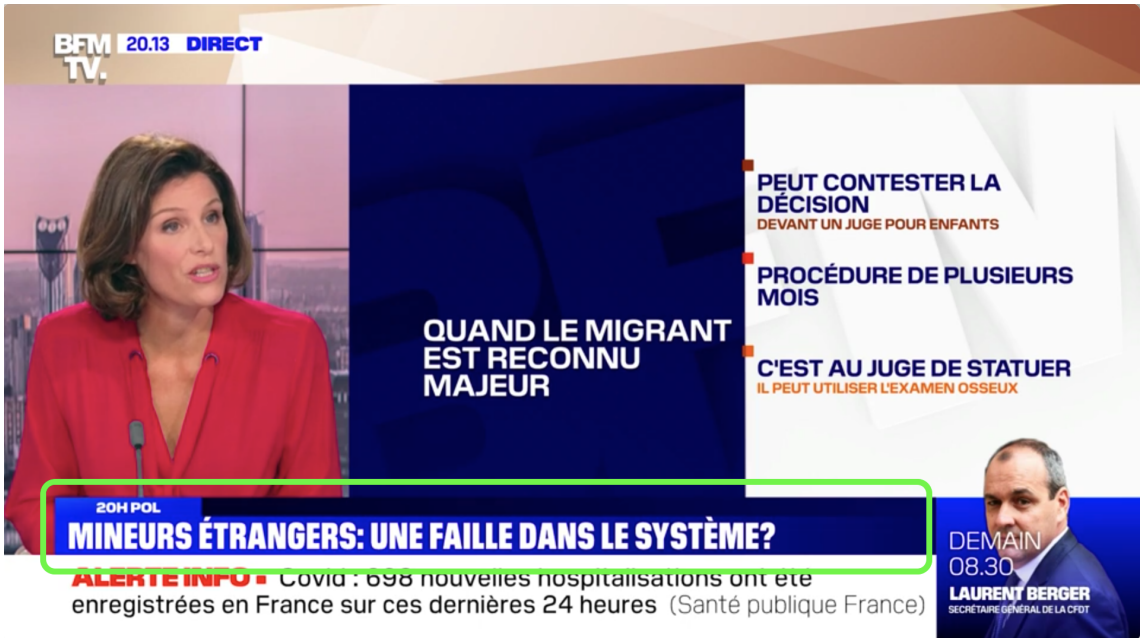
[Utiyama and Isahara, 2001] Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 499–506, Toulouse, France. Association for Computational Linguistics. – Cité page 22.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. – Cité page 23.

APPENDIX A

ANNEXE

A.1 Exemple de bandeau d'informations dans un journal télévisé résumant le sujet en cours de présentation



Le sujet en cours de présentation est entouré par le rectangle vert.

A.2 Un extrait de transcription du journal *Via-occitanie*

```
<SpeechSegment ch="1" sconf="1.00" stime="880.840" etime="883.390"
    spkid="S131" lang="fre" lconf="1.00" trs="1">
<Word id="2375" stime="881.77" dur="0.11" conf="0.73"> \`a </Word>
<Word id="2376" stime="882.07" dur="0.75" conf="0.89"> la </Word>
<Word id="2377" stime="882.84" dur="0.19" conf="0.99"> une </Word>
</SpeechSegment>
<SpeechSegment ch="1" sconf="1.00" stime="903.290" etime="909.640"
    spkid="S133" lang="fre" lconf="1.00" trs="1">
<Word id="2378" stime="903.66" dur="0.05" conf="0.42"> euh </Word>
</SpeechSegment>
```

```
<SpeechSegment ch="1" sconf="1.00" stime="909.640" etime="912.250"
    spkid="S135" lang="fre" lconf="1.00" trs="1">
<Word id="2379" stime="910.58" dur="0.35" conf="0.60"> y </Word>
</SpeechSegment>
<SpeechSegment ch="1" sconf="1.00" stime="937.380" etime="945.690"
    spkid="S140" lang="fre" lconf="1.00" trs="1">
<Word id="2380" stime="938.06" dur="1.81" conf="0.78"> \c ca </Word>
<Word id="2381" stime="940.67" dur="0.43" conf="0.67"> y </Word>
<Word id="2382" stime="944.68" dur="0.30" conf="0.52"> est </Word>
</SpeechSegment>
<SpeechSegment ch="1" sconf="1.00" stime="945.690" etime="947.680"
    spkid="S141" lang="fre" lconf="1.00" trs="1">
</SpeechSegment>
<SpeechSegment ch="1" sconf="1.00" stime="970.490" etime="973.770"
    spkid="S168" lang="fre" lconf="1.00" trs="1">
<Word id="2383" stime="972.89" dur="0.09" conf="0.29"> &lt;unk&gt;
    </Word>
</SpeechSegment>
<SpeechSegment ch="1" sconf="1.00" stime="974.130" etime="984.910"
    spkid="S144" lang="fre" lconf="1.00" trs="1">
<Word id="2384" stime="979.81" dur="0.09" conf="0.28"> et </Word>
</SpeechSegment>
```

<SpeechSegment> un groupe de souffle </SpeechSegment>
<Word> mot transcrit par le système </word>
id : id du mot
stime : temps début du mot
dur : la durée du mot
conf : score de confiance