Segmentation thématique de texte linéaire et non-supervisée :Détection active et passive des frontières thématiques en Français

Alexandre Labadié

Numéro d'identification :

Académie de Montpellier

Université Montpellier II

— **Sciences et Techniques du Languedoc** —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité	: **Informatique**
Formation Doctorale	: **Informatique**
École Doctorale	: **Information, Structures, Systèmes**

Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français.

par

Alexandre LABADIE

Soutenue le 3 Décembre Année universitaire 2008 - 2009

27ème Section : Informatique devant le Jury composé de :

Jacques Chaucké, Professeur, UMII, ... Président
Violaine Prince, Professeur, UMII, ... Directrice de thèse
Marc El-Bèze, Professeur, UAPV, ... Co-directeur de thèse
Brigitte Grau, Professeur, ENSIIE, ... Rapporteur
Bernard Levrat, Professeur, Université d'Angers, ... Rapporteur
Yves Bestgen, Professeur, Université catholique de Louvain, ... Rapporteur

Numéro d'identification :

Académie de Montpellier

Université Montpellier II

— **Sciences et Techniques du Languedoc** —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité	: **Informatique**
Formation Doctorale	: **Informatique**
École Doctorale	: **Information, Structures, Systèmes**

Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français.

par

Alexandre LABADIE

Soutenue le 3 Décembre Année universitaire 2008 - 2009

27ème Section : Informatique devant le Jury composé de :

Jacques Chaucké, Professeur, UMII, ... Président
Violaine Prince, Professeur, UMII, ... Directrice de thèse
Marc El-Bèze, Professeur, UAPV, ... Co-directeur de thèse
Brigitte Grau, Professeur, ENSIIE, ... Rapporteur
Bernard Levrat, Professeur, Université d'Angers, ... Rapporteur
Yves Bestgen, Professeur, Université catholique de Louvain, ... Rapporteur

Je dédie cette thèse à mes parents Marc et Marie-Christine, à ma petite soeur Claire et à Pascal. Ils m'ont toujours soutenu, même s'ils ne m'ont pas toujours compris.

Remerciements

Une thèse est un travail de longue haleine qui s'étend sur une longue période (trois ans pour moi) et même si c'est un travail personnel, c'est également un travail qui ne peut être accompli sans le soutient d'autres personnes. J'ai donc décidé de consacrer un peu de ce document lourd et ennuyeux qu'est une thèse à ces personnes qui ont participé, que ce soit par une intervention directe ou par leur soutient, à la réalisation de ma thèse. L'ordre dans lequel je remercie ces personnes n'a rien à voir avec une quelconque préférence ou importance, je remercie ici tous ceux qui m'ont aidé par ordre alphabétique et j'espère n'oublier personne.

Je remercie tout d'abord les membres de mon jury, **Jacques Chauché**, **Violaine Prince**, **Marc El-Bèze**, **Brigitte Grau**, **Bernard Levrat** et **Yves Bestgen** qui ont probablement contribué le plus directement à l'obtention de mon titre de docteur en informatique. Je m'attarderai sur certains d'entre eux plus loin.

Je remercie **Nicolas Béchet**, camarade de galère arrivé une année après moi. Sa sympathie et son ouverture d'esprit m'ont beaucoup aidées dans les différentes phases de la fin de ma thèse. Nicolas m'aura permis de ne pas m'enfermer totalement dans la solitude sur la fin de ma thèse (et m'aura ainsi probablement évité de finir neurasthénique).

Je remercie **Yves Bestgen**, professeur à l'université de Louvain en Belgique, d'avoir accepté de rapporter mon mémoire de thèse, mais aussi pour m'avoir fourni de précieux textes pour mes travaux. Je regrette seulement que nos échanges ce soient limités à des courriers électronique et qu'il n'est pas pu être présent en personne à ma soutenance. J'aurais aimé pouvoir échanger des points de vue avec lui en personne.

Je remercie **Jacques Chauché**, collègue de mon équipe et président de mon jury, avec qui j'ai collaboré durant mes travaux. Sa bonhommie contagieuse et sa sympathie en on fait un interlocuteur disponible et accessible dès le premier jour de ma thèse.

Je remercie **Julien Cotret**, vieil ami qui vient de nous rejoindre dans le cruel monde des thésards. Grâce à son paradoxal et si personnel mélange de décalage et de réalisme, Julien m'a apporté une vision du monde qui a surement eu de l'influence sur mes travaux.

Je remercie **Jurgen et Annan Darquenne**, des amis très proches. Ils ont toujours été là pour moi que ce soit pour me sortir de ma solitude ou pour s'amuser ensemble.

Je remercie **Marc El-Bèze**, mon codirecteur de thèse, qui malgré la distance et le manque de contacts entre nous aura toujours était disponible pour répondre à mes questions et me sortir de certaines impasses quand je lui demandais.

Je remercie **Brigitte Grau**, Professeur à l'ENSIIE, d'avoir accepté de rapporter ma thèse. Ses remarques constructives m'ont aidé à améliorer la qualité du document final.

Je remercie **Claire Labadié**, ma sœur et future institutrice. Sa gentillesse et la foi (parfois non justifiée selon moi) qu'elle a dans les capacités de son vaurien de grand frère,

m'ont toujours aidé dans les moments de doute.
Je remercie **Marc Labadié**, mon père, qui m'a toujours soutenue malgré son incompréhension manifeste (et non dissimulée) de mon travail. Il n'a pas hésité à me consacrer des journées entières pour m'aider à corriger les fautes d'orthographes dans ma thèse, et si cette dernière n'en contient pas trop c'est en grande partie grâce à lui.
Je remercie **Marie-Christine Labadié**, ma mère, qui a toujours su trouver les mots pour m'aider à garder les pieds sur terre. Malgré ses doutes sur la voie que j'ai choisie, elle n'a jamais cessé de me soutenir et de m'encourager.
Je remercie **Mathieu Laffourcade**, un collègue de l'équipe qui partage avec moi une certaine tendance au cynisme et à l'humour noir. Les discussions que nous avons eu sur de nombreux sujets (dont certains sérieux) ont été un excellent stimulant intellectuel.
Je remercie **Bernard Levrat**, Professeur à l'Université d'Angers, d'avoir accepté de rapporter ma thèse. Son exigence de rigueur m'aura aidé à corriger certains défauts de mon travail.
Je remercie **Violaine Prince**, ma directrice de thèse, qui m'a laissé une grande liberté d'action et de pensé dans mon travail, tout en restant toujours disponible. Elle a eu la patience de tolérer mes nombreux défauts pendant trois ans (et notamment mon manque de diplomatie dans mon écriture scientifique) et ses conseils avisés m'ont toujours permis de sortir des nombreuses impasses dans lesquelles mon tempérament parfois trop fonceur m'a envoyé. Je la remercie doublement car en dehors de son aide précieuse dans mon cheminement sur la difficile voie de la recherche, elle n'a pas hésité à sacrifier de son temps pour me décharger de certaines lourdeurs administratives (notamment sur la fin de ma thèse). Les échanges que nous avons eu, tant sur mon domaine de recherche, que sur des sujets comme la philosophie, la théologie ou encore la politique, ont toujours étaient particulièrement enrichissant.
Je remercie **Mathieu Roche**, un collègue de l'équipe qui m'a donné nombre de conseil avisés dans le domaine de la rédaction scientifiques notamment.
Je remercie **Jérémie Soullier**, mon plus vieil ami sur Montpellier. Sa franchise rafraichissante et son optimisme m'ont aidé à garder le moral durant les moments difficiles.
Je remercie **Mehdi Yousfi-Monod** qui me précédait de deux ans dans l'équipe. Mehdi à toujours était disponible quand j'avais besoin d'aide, il a joué le rôle de mentor en bien des occasion pour tout ce qui est d'être un bon thésard. Mehdi a également était un interlocuteur captivant, que ce soit dans nos discussion au sujet du monde du logiciel libre, dont il est un membre actif ou dans nos polémique sur la **LEM** [1].
Je remercie tout ceux que je n'ai pas cité ou que j'ai oublié, mais qui m'ont tout de même soutenu, aidé, supporté tout au long de ces trois ans.

1. Loi de l'Emmerdement Maximum, une généralisation du théorème de Murphy

Sommaire

Table des figures

Liste des tableaux

1

Introduction

Sommaire

De tout temps, le stockage de l'information et son exploitation ont posé problème. Depuis que l'homme sait fixer de l'information sur un support il cherche à en optimiser l'exploitation. Parmi les plus vieux exemples, on peut citer la bibliothèque d'Alexandrie. Preuve s'il en est que, déjà dans l'antiquité on se préoccupait déjà de la concentration, du stockage et de l'archivage de la connaissance afin d'en faciliter l'accès.

Avec l'avènement de l'informatique, l'humanité a (en grande partie) résolu le problème du stockage de grandes quantités d'informations. Il nous est maintenant possible de stocker au format numérique des milliards de documents dans un espace inférieur à celui d'un paquet de cigarettes. Si ces prouesses technologiques représentent sans conteste une avancée considérable pour nous tous, elles n'en présentent pas moins un inconvénient majeur : l'être humain ne peut utiliser une telle masse d'information sans des outils appropriés. En fait, nous sommes arrivés à un point où la trop grande quantité d'information disponible devient préjudiciable à son exploitation par l'homme. *« Trop d'information tue l'information »*.

Pour pallier ce problème, différentes solutions ont été trouvées, parmi elles la création de méthodes de structuration de l'information telles que les bases de données par exemple. Très efficaces pour structurer de l'information « artificielle » [2], ces méthodes montrent vite leurs limites face au langage naturel. En effet, chaque langue obéit à ses propres règles et suit ses propres structures (même si des similarités existent, notamment dans les langues proches). De plus, au sein de chaque langue existe une pléthore d'exceptions, rendant très difficile l'exploitation automatique des textes en langue naturelle. L'interprétation

2. Par opposition au langage naturel qui par essence est une forme « naturelle » d'information.

des messages tels qu'ils sont produits par un être humain est donc un problème complexe. Le domaine du Traitement Automatique des Langues Naturelles (TALN) se propose de chercher des solutions à ce problème d'automatisation de l'interprétation du langage naturel. Presque aussi vieux que l'informatique elle-même, le TALN peut retracer ses origines jusqu'en 1958 et les débuts de la guerre froide. C'est en effet à cette période que l'on a commencé à vouloir automatiquement traduire les millions de lignes de texte que comportaient les messages russes interceptés. La première des tâches à laquelle le TALN dut s'attaquer fut donc la traduction automatique, mais depuis cette époque beaucoup d'autres tâches sont venues s'ajouter à cet ancêtre historique. Ainsi le TALN regroupe (entre autres) des tâches aussi variées que : le résumé automatique, la fouille de texte, la correction orthographique, la génération automatique de texte, la reconnaissance et synthèse de la parole ou encore la reconnaissance de caractères manuscrits.
Nos travaux se situent dans le domaine de la fouille de texte, plus précisément autour de la tâche spécifique de la segmentation de documents textuels. Il faut bien différencier la segmentation **thématique** de texte et la segmentation de texte. La segmentation de texte regroupe en vérité bien des applications, depuis l'extraction de portions de texte lorsqu'elles sont mélangées avec de l'image, de la vidéo ou du son dans des documents multimédias ([Karatzas, 2003]), au regroupement de mots en morphèmes ou en unités linguistiques plus importantes, problème très présent dans les langues asiatiques basées sur des idéogrammes ([Wu & Tseng, 1993], [Yang & Li, 2005]). La segmentation **thématique** de texte n'est que l'une des nombreuses tâches diverses et variées que l'on regroupe maladroitement sous le nom de segmentation de texte. L'objectif d'une application de segmentation thématique de texte est de diviser un texte en plusieurs unités textuelles de plus petite taille, appelées segments thématiques ; chacun devant traiter d'un thème particulier aussi disjoint que possible des thèmes des segments précédents et suivants. Par la suite nous désignerons par segmentation thématique la tâche qui nous intéresse, à savoir la segmentation thématique de documents textuels.
La tâche de segmentation thématique ne présente que peu d'intérêt en elle-même, en dehors de fournir une « carte thématique » du texte. Par contre, elle étend assez sensiblement les possibilités d'autre tâches du TALN. En recherche d'information par exemple, la segmentation thématique peut servir à ramener une information plus précise et / ou plus concise ([Prince & Labadié, 2007]). En résumé automatique, elle peut permettre soit d'identifier plus aisément les passages les plus pertinents soit d'orienter le résumé sur un thème bien précis ([Barzilay & Elhadad, 2000], [McDonald & Chen, 2002], [Farzindar, 2004]). D'autres applications peuvent être imaginées comme l'indexation thématique de grands textes ou la génération automatique de tables des matières « thématiques ».

Si l'analyse de la structure du texte par la machine date de la fin des années 80, avec

par exemple la théorie sur la structure rhétorique ([Mann & Thompson, 1987]), la segmentation thématique n'a commencé à se détacher de la recherche d'information ou de la fouille de texte qu'il y a une dizaine d'années. Le développement de l'application TextTilling par Hearst ([Hearst, 1997]), ainsi que la thèse de Reynar ([Reynar, 1998]) marquent probablement le début de l'intérêt de la communauté du TALN pour cette tâche en tant que telle.

1.1 Problématique

Lorsque l'on s'intéresse à la segmentation thématique, on constate très rapidement que la problématique de la tâche est multiple. Il est couramment admis dans la littérature que la segmentation thématique a pour objectif de trouver au sein d'un texte des portions cohérentes thématiquement et distinctes des portions voisine ([Choi, 2000], [Georgescul *et al.*, 2006], [Lamprier *et al.*, 2008]), on appelle ces portions des **segments thématiques**. Mais avant même de rechercher ces segments thématiques dans un texte il faut se poser la question de la nature de ce que l'on entend par « thème ».

1.1.1 La notion de thème

Il existe beaucoup de définitions du mot « thème », chacune correspondant à un usage ou un domaine d'usage différent.
Le terme thème vient du grec *thema* qui signifie ce qui est proposé. Si l'on ouvre un dictionnaire, la définition que l'on lira sera : *« Sujet, idée sur lesquels porte une réflexion, un discours, une œuvre, autour desquels s'organise une action »* ([Larousse, 1997]). En linguistique le thème est : *« L'élément d'un énoncé qui est réputé connu par les participants à la communication »* ou encore *« Terme de la phrase (syntagme nominal) désignant l'être ou la chose dont on dit quelque chose. »* (on l'oppose souvent au rhème qui est l'information nouvelle apportée par l'énoncé). En musique la notion de thème est également présente et signifie : *« Fragment mélodique ou rythmique sur lequel est construite une œuvre musicale. »*. Nous passerons ici sur les thèmes astraux et militaires, pour nous concentrer sur ce que ces définitions ont en commun. Le thème est l'information centrale sur laquelle s'articule un acte de communication. Plus simplement, la définition que nous retiendrons de la notion thème dans nos travaux sera : *« Ce dont on parle »*, l'information principale communiquée par l'auteur. En cela nous nous rapprochons de l'étymologie du terme.
Dès lors, la notion de segment thématique devient moins floue et l'on comprend qu'un

segment thématique est une portion de texte qui « parle » d'un unique sujet. Il sera toutefois nécessaire de définir plus en détail ce qu'est un segment thématique et nous nous y attacherons dans cette thèse.

1.1.2 Le choix du non-supervisé

En TALN, nous pouvons distinguer deux manières différentes d'approcher un problème : Les approches supervisées et les approches non-supervisées.
Une approche est supervisée lorsqu'elle nécessite une phase d'apprentissage automatique où l'on cherche à générer les règles qui seront utilisées pour venir à bout de la tâche à partir d'une base de données d'apprentissage contenant des exemples déjà traités. En TALN, cette base de données d'apprentissage prend, en général, la forme d'un corpus de textes annotés pour la tâche. En segmentation thématique, elle prend la forme d'un corpus de textes dans lesquels les phrases frontières sont annotées. Quoi qu'il en soit, quelque soit le domaine du TALN dans lequel ce type de méthode est appliquée, la faille est toujours la même : La méthode est en général performante tant quelle est appliquée sur des données correspondant à ses données d'apprentissage. Dès que l'on sort de ce cadre, ses performances ont tendances à se dégrader.
Une approche « non-supervisée », quant à elle, s'appuie seulement sur l'information fournie par les données qu'elle doit traiter, et parfois, dans le TALN, sur des ressources linguistiques généralistes externes (thésaurus, dictionnaire, analyseur syntaxique, etc.). De telles approches sont plus généralistes et plus adaptables, mais rarement aussi efficaces qu'une méthode supervisée (lorsque cette dernière est appliquée à des données proches de ses données d'apprentissage).
Notre choix s'est donc porté sur les méthodes non-supervisées, ces dernières pouvant être appliquées à une plus vaste variété de textes. Nous développerons plus en détail les avantages et inconvénients des deux types d'approche dans notre chapitre 2.

1.2 Contribution de cette thèse

Un modèle de segmentation thématique non-supervisé basée sur une analyse en profondeur de la phrase et sur des informations stylistiques. Alors que bon nombre de méthodes non-supervisées s'appuient uniquement sur l'information lexicale pour en déduire la structure thématique du texte, notre approche exploite les informations présentes au niveau de la phrase. Sur la base de ces informations nous recherchons des indices stylistiques qui nous permettrons de reconstituer la structure thématique du

texte.

Une application de segmentation thématique modulaire et adaptable. Nous avons développé une application, nommée Transeg (tout comme notre approche), de segmentation thématique. Cette application permet notamment de tester tous les aspects du modèle et de faire varier tous ses paramètres.

Un protocole d'évaluation par l'homme des méthodes de segmentation thématiques. L'évaluation de la segmentation thématique est problématique. Comme de nombreuses tâches du TALN, c'est une tâche relativement subjective et donc peu adaptée à l'évaluation automatique. Nous avons donc développé un protocole d'évaluation manuel, basé sur des critères précis, plus adapté à la tâche. Les résultats récoltés lors de notre évaluation manuelle tendent à confirmer le caractère subjectif de la tâche.

1.3 Organisation de la thèse

Dans le chapitre 2 de cette thèse, nous nous attacherons à décrire les différentes approches existantes dans le domaine de la segmentation thématique. Nous y justifierons notre choix du non-supervisé en y décrivant brièvement les forces et faiblesses des méthodes supervisées. Puis nous nous intéresserons plus particulièrement aux méthodes non-supervisées, à leurs différences, mais surtout à leurs points communs. Nous essaierons ainsi de montrer qu'il existe des espaces encore non explorés dans ce domaine et qu'il pourrait être intéressant d'explorer.

Nous consacrerons le chapitre 3 à la présentation des bases théoriques sur lesquelles nous nous sommes appuyés pour développer notre approche. Dans une première partie nous y présenterons notre choix pour la représentation du texte : les vecteurs sémantiques. Puis nous nous attacherons à définir la notion de segment thématique. Enfin nous y étudierons les différentes possibilités qui se sont présentées à nous en matière de segmentation thématique active comme passive

Dans le chapitre 4, nous présenterons la mise en application des théories, principes et hypothèses que le chapitre 3 aura mis en avant. Nous y décrirons le développement de l'application qui par la suite nous servira pour tester notre cadre théorique. La première partie de ce chapitre servira à décrire globalement l'architecture de l'application. Puis nous expliquerons comment sont générés les vecteurs sémantiques évoqués dans le chapitre 3. Nous consacrerons la partie suivante aux différents outils de comparaison vectorielle que nous avions à notre disposition et présenterons la distance de concordance que nous avons développée spécifiquement pour la tâche de segmentation thématique. La dernière partie de ce chapitre présentera certains points spécifiques du développement de l'application

qui présentaient un intérêt.
Nous dédierons le chapitre 5 à la description de notre cadre expérimental, ainsi qu'aux résultats que nous avons obtenus. D'abord nous nous attacherons à identifier les difficultés propres à l'évaluation d'une tâche de TALN et plus spécifiquement de segmentation thématique. Nous consacrerons la partie suivante à décrire notre corpus d'expérimentation composé en réalité de deux corpus. La troisième partie de notre chapitre d'évaluation présentera c99, l'algorithme dont nous nous servons comme base de comparaison lors de nos expériences. Ensuite, nous présenterons et commenterons les résultats de notre approche et de ses variantes sur le corpus. La dernière partie sera consacrée à notre évaluation humaine, nous y présenterons le protocole que nous avons mis en place et les résultats que nous avons obtenus.
Nous conclurons dans le chapitre 6 avec une synthèse du travail présenté dans cette thèse et des perspectives d'évolution pour nos travaux.

2

Segmentation thématique de texte : panorama et état de l'art

Sommaire

2.1 Introduction

Le domaine de la segmentation thématique est un domaine qui a donné lieu à de nombreux travaux ces dernières années. Les applications directes de cette tâche sont, certes, peu nombreuses (comme la création de table des matières thématiques par exemple), mais ses applications indirectes sont, elles, bien plus nombreuses. La segmentation thématique peut par exemple être utilisée pour améliorer les performances de systèmes de question-réponse ([Prince & Labadié, 2007]) en fournissant des portions de texte thématiquement proches de la question. Le résumé automatique peut également être amélioré, soit en permettant un résumé thème à thème, soit en proposant un résumé thématique. Plus généralement, de nombreuse tâches du TALN peuvent bénéficier de la segmentation thématique.

Bien que nous ayons choisi d'opter pour une approche non-supervisée, ce panorama a pour objectif de présenter ce qui ce fait dans le domaine de la segmentation thématique, que ce soit de manière supervisée ou non-supervisée. Nous nous intéresserons particulièrement à l'aspect passif ou actif de ces méthodes. Les méthodes « passives » ne cherchent pas

à retrouver les frontières thématiques d'un texte, mais à regrouper les phrases en segment. Les frontières apparaissant ainsi par « défaut », d'où notre dénomination de méthodes à détection passive des frontières. Les méthodes « actives », au contraire, tentent d'identifier spécifiquement les propriétés des frontières thématiques, les segments thématiques devenant l'espace qu'il y a entre deux frontières, nous les appelons méthodes à détection active des frontières.
Nous consacrerons donc notre section 2.2 à définir plus en détail les concepts d'approches actives et passives. Nous y verrons notamment qu'il s'agit là plus d'une classification vis à vis de la démarche intellectuelle qui motive ces approches que d'une classification sur les moyens (algorithmes, ressources, etc.) utilisés par ces approches.
Dans la section 2.3, nous nous attacherons à décrire les approches que nous avons catégorisées comme passives. Nous nous concentrerons sur les plus emblématique et sur leurs points communs et leurs différences.
Nous verrons dans la section 2.4 les approches que nous considérons comme actives. Une fois de plus nous nous focaliserons sur les plus emblématique en nous intéressant à ce qui qui les rassemble et ce qui les différencie.
Nous ferons un comparatif général de ces différentes méthodes dans la section 2.5. Dans ce comparatif nous verrons que ces méthodes s'appuient toutes sur un principe commun et n'utilisent pas ou peu toutes une catégorie d'informations qui peuvent pourtant apporter beaucoup à la tâche.
Nous conclurons ce chapitre dans la section 2.6.

2.2 Segmentation passive et active : définition des concepts

Lorsque l'on doit catégoriser des méthodes en TALN, on applique souvent le schéma classique de l'opposition supervisé contre non-supervisé. Cette catégorisation se base sur une différence méthodologique entre les approches et on peu la rapprocher des travaux de [Ferret, 2006] qui défini deux grandes catégories d'approches : celles exploitant des connaissances externes aux texte (les approches exogènes) et celles n'exploitant que le contenu du texte (les approches endogènes).
Bien que satisfaisante, cette vision de la segmentation thématique ne s'intéresse qu'à la manière dont on procède, et non à l'objectif que c'est fixé le créateur d'une méthode. Nous avons choisi de nous intéresser de plus près à cet objectif.
Notre distinction entre segmentation active et passive vient donc de cette volonté de catégoriser les méthodes de segmentation thématique, non pas sur la base de leur métho-

dologie, mais sur la base de la motivation initiale, l'objectif de leur créateur. Au delà de vouloir segmenter thématiquement le texte, on peut isoler deux visions de la segmentation thématique :

- Il y a les méthodes qui cherchent à regrouper les phrases similaires entre elles, formant ainsi des segments thématique. Dans ces méthodes, les frontières apparaissent comme une conséquence de ce regroupement, elles sont donc retrouvées « par défaut », passivement. Ce sont les méthodes que nous appelons « à détection passive » des frontières.
- Il y a les méthodes qui cherchent à identifier les propriétés des frontières pour localiser ces dernières. Dans ces dernières, les frontières thématiques n'apparaissent pas par défaut, mais sont activement recherchées. D'où notre dénomination de « détection active » des frontières.

Bien entendu, et nous allons le voir par la suite, il y a un lien entre la méthodologie employée et la démarche intellectuelle qui est à l'origine du choix de la méthodologie. Mais, si l'on peut supposer sans trop se tromper que les approches endogènes sont passives et que les approches exogènes sont actives, nous verront que ce n'est pas toujours le cas.

2.3 Les approches passives

Comme nous l'avons défini dans l'introduction, les méthodes passives cherchent à regrouper les phrases en segments thématique et déduisent de ce regroupement la position des frontières.

Parmi les plus représentative des approches passives, mais aussi les plus originales des méthodes de segmentation thématique, on citera toute un groupe que nous appellerons : Méthodes graphiques.

2.3.1 Les méthodes graphiques

En passant par une représentation graphique des termes, il est plus facile de visualiser leur répartition le long du document étudié. Ainsi la méthode du nuage de points, présentée par [Helfman, 1994] emploie cette représentation pour la recherche d'information. Le principe est de positionner sur un graphique chaque occurrence des termes du document. Dans cette représentation, un terme apparaissant à une position i et une position j du texte, sera représenté par les 4 couples (i, i), (i, j), (j, i) et (j, j). Les portions du document où les répétitions de termes sont nombreuses apparaîtront alors sur le graphique comme les zones de forte concentration de points.

Cette approche visuelle de la représentation d'un texte a été reprise et adaptée à la segmentation thématique par [Reynar, 1998] dans son algorithme *DotPlotting*. L'idée est d'identifier les segments thématiquement cohérents sur le graphique en cherchant les limites des zones les plus denses. La densité d'une région du graphique est calculée en divisant le nombre de points présents dans la région par l'aire de cette dernière. L'objectif de *DotPlotting* est d'isoler les segments thématiques soit en maximisant leur densité, soit en minimisant la taille des zones « vides » entre les segments. On notera que, dans son principe, cette méthode est très proche des méthodes utilisant des matrices de similarité présentées plus loin.
Cette approche a même inspiré des méthodes originales, comme celle proposée par [Ji & Zha, 2003], qui consiste à remplacer le problème de segmentation thématique par un problème de segmentation d'image. Cette méthode utilise une technique de diffusion anisotropique sur la représentation graphique de la matrice de distance afin de renforcer les contrastes entre les zones denses et les frontières.
La transformation du texte en un ensemble de points, pour ensuite retrouver les segments thématiques en regroupant ces points en nuages correspond totalement à la description que nous avons faite des méthodes à détection passive. On regrettera que ces méthodes aient une approche très peu « linguistique » de la segmentation thématique. En effet, en réduisant le texte à une représentation graphique de ses termes, on perd toute notion de compréhension et on se contente alors de compter des mots.

Plus traditionnelles que les méthodes graphiques et surtout plus répandues, les approches s'appuyant sur un calcul de distance ou de similarité sont surement parmi les plus populaires en segmentation thématique.

2.3.2 Les méthodes par calcul de distance / similarité

Les méthodes de segmentation à base de similarité ou de distance considèrent les différentes portions de texte du document à traiter comme autant de vecteurs. Les composantes des vecteurs étant, dans la plupart des cas, les fréquences d'apparition des mots au sein de la portion de texte, après que celle-ci ait été débarrassée des mots inutiles (mots jugés comme peu porteurs de sens). Parfois, cette fréquence des mots est pondérée par un IDF (Inverse Document Frequency), pour renforcer l'importance des mots supposés thématiquement saillants.
L'objectif de ces méthodes est donc de mesurer la proximité ou l'éloignement des portions de textes étudiées grâce à l'angle que forment leurs vecteurs représentatifs. Elles s'appuient donc en général sur le cosinus de cet angle, qu'elles considèrent comme la similarité. La

similarité est ensuite exploitée de diverses manières.
L'algorithme c99 ([Choi, 2000]) est un bon exemple de ces approches basées sur la similarité entre portions de textes. Cet algorithme s'appuie sur une matrice de similarité entre phrases pour ensuite construire une matrice de rang, correspondant à un classement local des similarités. C'est cette matrice de rang qui servira à retrouver les segments thématiques [3]. On notera que rapidement c99 a été amélioré pour utiliser un peu d'information sémantique en intégrant LSA [4] ([Choi *et al.*, 2001]).
Le clustering est également une approche possible de la segmentation thématique par calcul de similarité. En effet, regrouper des phrases en segments thématiques est assez similaire à la tâche consistant à regrouper des éléments d'un ensemble en classes. La seule contrainte supplémentaire est que les phrases d'un segment thématique doivent se suivre dans le texte. Ainsi, *ClassStruggle* ([Lamprier *et al.*, 2008]) utilise un algorithme de clustering évolutif pour segmenter thématiquement des textes.
Les méthodes par calcul de distance ou de similarité sont typiquement des méthodes à détection passive des frontières. Mais, nous le verrons dans la section suivante, certaines approches utilisant des distances ou des similarités peuvent être classées dans les approches active.

Au final, nous pouvons constater que les approches passive sont toutes basées sur la notion de cohésion lexicale telle que la définie [Morris & Hirst, 1991] et n'exploitent pas, ou très peu d'autres informations telle que la syntaxe de la phrase ou certain indice stylistique.

2.4 Les approches actives

Les approches actives cherchent à identifier les propriétés des phrases frontières pour les localiser. Cette démarche de recherche volontaire des frontières a bien entendu eu une grande influence sur les méthodologies employées. Aussi, peut on s'attendre à une majorité de méthodes utilisant des ressources externes au textes pour définir les propriétés des frontières thématiques (des approches exogènes selon la définition de [Ferret, 2006]). Toutefois, nous allons voir que certaines méthodes tentent de retrouver les frontières thématiques, sans s'appuyer sur des données externes (et sont donc endogènes toujours selon [Ferret, 2006]).

Lorsque l'on cherchent à identifier un type de phrase particulier, comme les phrases fron-

3. L'algorithme c99 sera décrit en détail dans le chapitre 5.
4. Latent Semantic Analysis ou Analyse Sémantique Latente.

tières par exemple, le premier réflexe que l'on a est d'apprendre à quoi ressemble ces phrases sur un corpus pour construire un modèle, puis d'utiliser ce modèle pour les identifier.

2.4.1 Les méthodes supervisées

Les méthodes supervisées s'appuient sur un apprentissage utilisant une base de données d'apprentissage pour en déduire des règles. En segmentation thématique, ces méthodes utilisent un corpus d'apprentissage composé de textes dont les frontières thématiques ont été annotées. En utilisant ces annotations les méthodes supervisées construisent un modèle permettant de retrouver les frontières thématiques d'autres textes. Les principales différences entre ces méthodes viennent donc de ce qu'elles apprennent et de comment elles l'apprennent.

Ainsi, [Georgescul *et al.*, 2006] proposent d'utiliser des machines à vecteurs de support (SVM) pour apprendre une distribution de mots correspondant aux phrases frontières et une distribution correspondant aux phrases qui ne sont pas des frontières. A partir de ces deux modèles, la méthode donne à chaque phrase du texte une probabilité d'appartenir à chacune des catégories définies par le modèle, puis choisit, en fonction de ces probabilités, qu'elles sont les phrases frontières.

Dans la même ligne, [Reynar, 2002] construit un modèle qui permet de dire si une « région de texte[5] » traite le même thème qu'une autre région.

D'autres ne s'intéressent qu'à des mots indicateurs, des « marqueurs linguistiques », et construisent des modèles statistiques pour identifier les phrases frontières ([Beeferman *et al.*, 1997, Beeferman *et al.*, 1999]).

Dans la ligné des marqueurs linguistiques, on citera les travaux de [Passonneau & Litman, 1993, Passonneau & Litman, 1997] qui se sont attachés à construire des modèles pour trois grandes catégories d'indice linguistiques que sont les propositions nominales de références, les mots indicateurs et les pauses. Leur travail portant sur la segmentation de discours oraux, la notion de pause dans le discours, qui nous est inconnue dans un cadre purement textuelle, est importante. En dehors de cette particularité, ils construisent des modèles pour chacun de ces indices linguistiques en ce basant sur des textes segmentés par des juges humains puis utilisent ces modèles pour retrouver les frontières thématiques. Leur originalité vient du fait que chacune des catégories de marqueurs linguistiques est traitées différemment. Leur travaux utilisant beaucoup d'information prosodique, ils sont difficilement applicable à un cadre textuelle n'incluant pas cette information.

5. Un groupe de 230 mots, ce nombre étant la taille moyenne des segments thématiques du corpus d'apprentissage.

Le problème de ces méthodes, c'est qu'elles héritent du travers reconnu des méthodes supervisées en général (quelque soit le domaine) : performantes sur des textes proche de leur corpus d'apprentissage, elles deviennent inefficaces dès qu'elles sortent de ce cadre. De plus les phases d'apprentissages peuvent parfois être lourdes et couteuses, surtout si peu de ressources sont disponibles pour l'apprentissage comme c'est souvent le cas en français.

Certaines approches combinent des ressources externes ou un apprentissage supervisé et une méthode endogène pour segmenter thématiquement. Nous appelons c'est méthodes « hybrides »

2.4.2 Les méthodes hybrides

Parmi les méthodes hybrides, on citera [Ferret *et al.*, 2001] qui combinent une ressources externes, des marqueurs linguistiques génériques et un modèle statistique plus « classique » pour retrouver les frontières thématiques du texte. Cette approche utilise des expressions qui marquent le début des *cadres* thématique ([Charolles, 1997]) avec soit une analyse typiquement endogène comme celle proposée par [Hearst, 1997] (décrite plus loin), soit une méthode utilisant des ressources externes pour enrichir le texte et généraliser la notion de cohésion lexical à une notion plus étendue de cohésion du champ lexical. La force de cette approche est que les marqueurs ne sont pas juste identifiés par leur lexie, mais également grâce à un ensemble de règles (sept précisément) décrivant les condition dans lesquelles le marqueur doit apparaître.
Plus récemment, [HP2006], vainqueurs de DEFT'06 ([Azé *et al.*, 2006] [6]), ont obtenu de bons résultats en combinant un apprentissage des marqueurs linguistiques par un modèle $n-gram$ avec le segmenteur probabiliste de [Utiyama & Isahara, 2001]. Cette approche hybride a malheureusement les défauts propres aux méthodes supervisées : en dehors du cadre du corpus d'apprentissage ces performances sont celles du segmenteur probabiliste, voir moins bonnes. En effet, les données issues de l'apprentissages peuvent être source de bruit et donc dégrader les résultat du segmenteur probabiliste.
Cette combinaison entre l'endogène et l'exogène donne en général de bon résultats, mais pose toujours le problèmes des ressources externes. Même lorsqu'elle ne sont pas « apprisses » [7], comme dans le cas de [Ferret *et al.*, 2001], elle doivent être « acquises », ce qui peut ce révéler lourd et contraignant.

6. Le Défi Fouille de Textes est un défi francophone consacré à la fouille de texte, l'édition 2006 portait sur la segmentation thématique.
7. Dans le sens d'un apprentissage automatique supervisé ou non tel que nous le rencontrons souvent dans le domaine du TALN

Parmi les méthodes endogènes actives, les plus courantes sont probablement les méthodes à base de chaînes lexicales.

2.4.3 Les méthodes par chaînes lexicales

Le principe des méthodes par chaînes lexicales est relativement simple. Il consiste à relier les occurrences d'un même terme, tout au long du texte, entre elles pour former une chaîne. Chaque terme du texte forme donc d'une chaîne et l'utilisateur détermine à partir de quelle distance entre deux occurrences l'algorithme considère la chaîne comme rompue [8]. Lorsqu'une phrase se trouve être le point de départ d'un grand nombre de chaînes (une fois encore ce paramètre doit être déterminé), alors elle est considérée comme une frontière thématique.
Ainsi, la méthode *Segmenter* présentée par [Kan *et al.*, 1998], procède selon ce principe pour effectuer une segmentation thématique du document étudié. *Segmenter* rajoute toutefois une subtilité, à savoir que la catégorie syntaxique du terme formant la chaîne entre en compte dans le calcul de la distance à partir de laquelle l'algorithme considère qu'il y a rupture. Cette originalité mérite d'être notée dans la mesure où peu de méthodes exploitent autre chose que le mot lui-même. L'introduction de la catégorie syntaxique dans les paramètres de l'algorithme suppose une certaine analyse de la phrase. Même si cette analyse reste très superficielle, elle apporte tout de même un complément d'information non négligeable. En modulant la distance nécessaire pour rompre une chaîne en fonction de la catégorie syntaxique du mot, *Segmenter* peut donner plus d'importance aux catégories syntaxiques très porteuses de sens (comme les verbes ou les noms) et moins à des catégories ayant moins de « poids sémantique » (comme les adjectifs par exemple [9]).
SeLECT développée par [Stokes *et al.*, 2004] est très similaire à *Segmenter* et s'appuie intégralement sur le concept de chaînes lexicales.
Text Tilling est une également approche utilisant les chaînes lexicales, mais qui utilise aussi d'autres critères. Proposée par Hearst ([Hearst, 1993], [Hearst, 1994], [Hearst, 1997]), elle utilise un score de cohésion qui est attribué à chacun des blocs de texte en fonction du bloc qui le suit. Le score de chaque bloc étant lui-même calculé à partir d'un score dit « lexical » attribué à chaque paire de phrases en fonction de la paire qui la suit. Ce score lexical dépend de paramètres tels que le nombre de mots en commun, de mots nouveaux et de chaînes lexicales actives dans les phrases considérées. Le score de chaque segment de texte est alors le produit scalaire normalisé des scores de chacune des paires de phrases

8. Cette distance peut être en nombre de phrases, de mots, de propositions, etc.
9. Il faudrait toutefois préciser que la catégorie syntaxique d'un terme n'est pas un indice suffisant pour déterminer le « poids sémantique » d'un mot. Sa fonction au sein de la phrase est également utile.

qu'il contient. Si un segment présente un score très différent des segments précédents et suivants, alors la rupture thématique se situe au sein de ce segment. *Text Tilling* est donc une à la fois une méthode utilisant un calcul de similarité et une méthode utilisant les chaînes lexicales. Si l'algorithme original présenté dans [Hearst, 1993] n'incluait pas de chaînes lexicales, elle les a inclues dans le calcul de son score de cohésion assez rapidement ([Hearst, 1994]).
Enfin, on pourra citer les travaux de [Sitbon, 2004] qui ont la double originalité de fusionner des approches à base de chaînes lexicales et de similarité et de ne pas considérer seulement les états « rompue » ou « active » pour chaque chaîne lexicale, mais plutôt un niveau d'activité d'une chaîne donnée à un moment donné du texte. La méthode enrichit la matrice de similarité que l'algorithme c99 [10] utilise.
La segmentation par chaînes lexicales est clairement une approche par détection active des frontières. Simple à comprendre et à mettre en œuvre, ces méthodes ont tout de même le défaut de ne s'appuyer que sur une information lexicale de base, parfois agrémentée d'information syntaxique (comme pour [Kan *et al.*, 1998]).

Peu courante en détection active, les méthodes par calcul de similarité ou de distance ne sont toutefois pas absentes.

2.4.4 Les méthodes par calcul de distance / similarité (encore)

Comme nous l'avons déjà dit dans la section précédente, ces méthodes ce basent sur un calcul de distance ou de similarité entre différente portion de textes. Si les méthodes passives utilisaient surtout la similarité, les méthodes actives se basent plutôt sur des distances. Cette différence peut paraître anodine, mais est représentative de la différence entre approches actives et passives.
Ainsi, *SegGen*, une méthode originale proposée par [Lamprier *et al.*, 2007], considère la segmentation thématique comme une tâche d'optimisation multi-critères et utilise un algorithme génétique pour optimiser les deux critères que sont la cohérence interne des segments thématiques et la dissimilarité avec les segments adjacents. Cette approche originale a le double mérite d'identifier les propriétés essentielles d'un segment thématique telles que nous les définirons dans le chapitre 3 et, en un sens, de chercher à réunir les notions de détection active et passive.
On notera également l'approche inhabituelle de [Utiyama & Isahara, 2001], qui construit au fil du texte un modèle local de distribution des mots et considère qu'il y a rupture thématique lorsque ce modèle change. Cette approche part du principe que deux thèmes

10. L'algorithme c99 sera présenté plus en détail dans le chapitre 5

différents devraient avoir deux distributions différentes.
Plus classique, la première version de *Text Tilling* ([Hearst, 1993]), n'incluait pas de chaînes lexicales et ne se basait donc que sur une similarité entre blocs de texte pour repérer les frontières thématiques.
Dans la même ligne, [Brants *et al.*, 2002] recherchent une baisse de similarité entre blocs de texte pour déterminer les frontières thématiques. Leur originalité est de se baser sur une représentation PLSA [11] du texte qui intègre notamment une notion de classes latentes. Cette approche combine donc un apprentissage avec une approche plus classique de calcul de distances / similarités, puisque lors de l'entrainement de PLSA les phrases frontières sont identifiées (entraînant donc probablement l'émergence d'une classe latente « frontière »).
Également basés sur un calcul de distance entre blocs les travaux de [Kaufmann, 1999] tirent leur originalité de l'usage d'une représentation du texte de type « *WordSpace* » mais intégrant une notion de contexte habituel pour le mot. Sur la base d'un corpus [Kaufmann, 1999] apprend les collocations usuelles d'un mots et construit une représentation qui les intègre. Lors des calculs de distance entre blocs, ne seront pas seulement pris en compte les mots, mais aussi leurs contextes habituels. On remarquera que ce type de représentation se rapproche énormément de LSA.

Une fois encore, nous constatons que l'information, même lorsqu'elle est extérieure au texte, reste très près du mot et ne se détache que rarement du simple niveau lexical.

2.5 Comparatif des différentes approches

Bien entendu, nous n'avons pas présenté ici tous les algorithmes et toutes les méthodes de segmentation thématique existants. Nous avons toutefois fait un tour complet des différents types d'approches qu'il est possible de trouver en segmentation thématique.
La table 2.1 présente un résumé des propriétés de différentes méthodes que nous avons vues dans ce chapitre. On observe que peu de méthodes exploitent l'information syntaxique ou sémantique et aucune l'information stylistique. Même lorsque les méthodes essayent d'intégrer un peu plus que seulement de l'information lexicale, elles ne le font que très partiellement.
Lorsque *Segmenter* utilise de l'information syntaxique ce n'est que pour récupérer la catégorie syntaxique des mots. Lorsque c99 intègre de l'information sémantique, c'est au travers d'une analyse sémantique latente qui n'est autre qu'une décomposition en valeurs

11. Probabilistic Latent Semantic Analysis

	Active	Passive	Lex. coh.	End./Exo.	Syntax.Syntaxique	Sém.Sémantique	Style
Segmenter	oui	non	oui	Endo.	part.	non	non
SeLECT	oui	non	oui	Endo.	non	non	non
TextTilling	oui	non	oui	Endo.	non	non	non
[Sitbon, 2004]	oui	oui	oui	Endo.	non	non	non
DotPlotting	non	oui	oui	oui	Endo.	non	non
c99	non	oui	oui	Endo.	non	part.	non
ClassStruggle	non	oui	oui	Endo.	non	non	non
SegGen	oui	oui	oui	Endo.	non	non	non
[Utiyama & Isahara, 2001]	oui	non	oui	Endo.	non	non	non
[Ferret *et al.*, 2001]	oui	non	oui	Les deux	part.	non	non
[HP2006]	oui	non	oui	Les deux	non	non	non
[Georgescul *et al.*, 2006]	oui	non	oui	Exo.	non	non	non
[Reynar, 2002]	oui	non	oui	Exo.	non	non	non
[Beeferman *et al.*, 1997]	oui	non	oui	Exo.	non	non	non

TABLE 2.1 – Les algorithmes de segmentation thématique

singulières sur une matrice d'occurrences [12] de mots.

De plus, si nous regardons le fonctionnement des différentes méthodes présentées ici, nous nous rendons rapidement compte que ces dernières se ressemblent énormément. Cette ressemblance est très frappante lorsque l'on s'intéresse aux méthodes graphiques et à celle basées sur des calculs de similarité ou de distance. Difficile de ne pas voir une forte similitude entre *DotPlotting* et c99. Si l'une utilise une représentation graphique, et l'autre une matrice de rang, au final les deux cherchent des zones denses dans leur représentation. Même les méthodes utilisant des ressources externes au texte semble ce limiter à n'utiliser que des distributions de mots. [HP2006] se basent sur des $n-gram$ appris sur un corpus par exemple. Et si [Ferret *et al.*, 2001] utilisent des ressources linguistiques génériques et un ensemble de règles contextuelles pour identifier les frontières, celles-ci tournent autour de marqueurs typique de début de *cadres* thématiques, donc des mots.

Au final, nous constatons que toutes les méthodes présentées ici utilisent le mot pour identifier une structure se situant au niveau du texte. Si l'information lexicale et la notion de cohésion lexicale telle qu'elle est définie par [Morris & Hirst, 1991] est incontestablement cruciale dans la recherche de la structure thématique d'un texte, ignorer l'information portée par la phrase (syntaxique ou sémantique) et par des portions entières de texte (stylistique) revient à se priver d'indices probablement aussi importants que ceux apportés par les mots.

12. Nous décrirons plus en détail LSA dans le chapitre 3.

2.6 Conclusion

Dans la section 2.2, nous avons précisé les notions de détection active et passive évoquées dans le titre de cette thèse. Nous y avons notamment insisté sur le fait que cette distinction s'attache surtout à la démarche et pas à la méthode employée.
Dans la section 2.3, nous avons passé en revue quelques approches de segmentation thématiques que nous avons qualifiées de passive. Nous avons pu constater que ce sont des approches à détection passive se basent presque exclusivement sur la cohésion lexical et n'exploite que rarement d'autre piste (on notera l'usage de LSA pour [Choi *et al.*, 2001]).
Dans la section 2.4, nous avons présenté les approches de segmentation thématique actives. Nous avons pu constater que toutes les méthodes exogène entre dans cette catégorie, mais également que les méthodes actives souffraient du même travers que les méthodes passives : la sous utilisation d'information autre que lexicale.
Nous avons enfin consacré la section 2.5 à une comparaison des approches différentes approches (actives et passives). Cette comparaison nous à permis de faire ressortir d'autant plus la racine commune de toute ces méthodes : le mot. De fait, nous avons pu constater qu'il existe un vide

Ce bref survol de ce qui se fait en segmentation thématique à l'heure actuelle nous a permis de voir qu'il existe des pistes inexplorées dans ce domaine. La phrase est une construction signifiante. En effet, le sens d'une phrase n'est généralement pas le résultat de la somme des sens des mots qui la composent. Dans ces conditions ignorer l'information portée par la phrase revient à gâcher une ressource qui devient de plus en plus exploitable. Si dans les débuts du TALN, il n'existait pas ou peu d'outils permettant l'exploitation correcte de cette ressource, ce n'est plus le cas aujourd'hui. Il existe maintenant des analyseurs de plus en plus performants, comme SYGFRAN ([Chauché, 1984]) que nous utilisons dans nos travaux [13] ou encore TreeTagger ([Schmid, 1994]) qui est très utilisé dans la communauté. Certes, ces outils restent très perfectibles et sont encore loin de fournir des résultats parfaits, mais ils s'améliorent petit à petit et commencent à offrir des possibilités qu'il serait dommage d'ignorer.

13. Nous décrirons plus en détail le fonctionnement de SYGFRAN dans le chapitre 4

3

Détection de changement de thème et optimisation de la cohérence thématique

Sommaire

3.1 Introduction

Cette section présente les bases de notre réflexion sur la segmentation thématique, ainsi que les outils théoriques que nous avons utilisés pour mener à bien cette réflexion.
La section 3.2 expose les différentes représentations du texte que nous avons envisagées, et décrit plus en détail celle que nous avons choisi d'employer.
La section 3.3 s'attache à définir la notion de segment thématique qui est au cœur de notre réflexion. Elle est suivie d'une courte section 3.4 où nous présentons les notations utilisées dans ce chapitre.
La section 3.5 va présenter notre réflexion théorique sur la segmentation thématique par détection active des frontières thématiques, tandis que la section 3.6 s'intéressera à la

détection passive.
Nous conclurons se chapitre dans la section 3.7.

3.2 La représentation du texte par vecteur sémantique de phrase

Si l'on considère la langue comme une forme structurée et formalisée de la communication humaine, alors la langue écrite, le texte, est une représentation encore plus formalisée et structurée de la langue. Malgré cela, la langue écrite reste un phénomène humain, généré par l'humain, à destination de l'humain, donc a priori peu ou pas vraiment inadapté au traitement par un ordinateur.
Afin de faciliter le traitement du texte par la machine, il est indispensable de disposer d'une représentation en adéquation avec la tâche à accomplir. Cette représentation doit simplifier le traitement du texte, tout en conservant un maximum d'informations utiles à notre objectif de segmentation thématique.
Cette section présente la démarche qui nous a conduit à choisir une représentation par vecteur sémantique de phrase pour nos travaux sur la segmentation thématique. Nous présentons d'abord le choix de notre unité atomique, puis le mode de représentation à proprement parler.

3.2.1 Échelle de représentation : une unité atomique pour la segmentation thématique

Avant même de choisir un mode de représentation du texte, il nous faut choisir l'échelle de la représentation, la taille de l'élément plus petit et du plus basique que nous utiliserons.
Notre tâche consistant à découper le texte en plusieurs segments, cela nous interdit d'office l'usage du texte comme unité atomique. Ce qui est possible dans une tâche consistant à comparer des textes entre eux comme la classification de texte par exemple ([Chauché & Prince, 2007]), ne l'est pas pour une tâche consistant à analyser un texte en détail comme la segmentation thématique.
Pour la même raison, le choix du paragraphe comme unité atomique paraît inapproprié. En effet, le résultat attendu d'une application de segmentation thématique est un ensemble de portions de texte (les segments thématiques) longues de plusieurs phrases. Ces

portions peuvent facilement être assimilées à des paragraphes [14] et donc le paragraphe ne convient pas comme unité atomique.
Notre choix s'est donc porté sur la phrase comme unité atomique, et ce principalement pour quatre raisons :

- La phrase est une unité syntaxique qui obéit à des règles de bonne formation qui assurent sa complétude et sa cohérence. Idéalement, la syntaxe ne dépasse pas le cadre de la phrase. La phrase se suffit donc à elle-même sur le plan syntaxique.
- La phrase possède une unité sémantique (ou unité de communication), c'est-à-dire, un contenu transmis par le message (sens, signification...). Ce qui signifie que la phrase se suffit à elle-même du point de vue sémantique [15], ce qui n'est pas le cas d'unités syntaxiques plus petites telles que la proposition ou le mot. En effet, selon Jakobson ([Jakobson, 1963]), le mot seul n'est rien, il ne se définit que par rapport aux autres éléments de la phrase. La proposition est, quant à elle, difficilement repérable et exploitable. La phrase apparaît donc comme étant le plus petit élément possédant une unité sémantique à l'échelle du texte, un élément atomique.
- La phrase est facilement identifiable. Diviser le texte en une succession de phrases est réalisable en se basant sur quelques heuristiques simples (ponctuation, formes particulières, etc.) et ce avec une fiabilité plus que satisfaisante [16].
- La phrase est l'unité atomique « traditionnelle » de la segmentation thématique. Les corpus de références et les différentes méthodes déjà existantes ([Stokes *et al.*, 2004], [Lamprier *et al.*, 2007],[Hearst, 1997], [Choi, 2000] sont des exemples parmi d'autres).

3.2.2 Choix d'une méthode de représentation : la représentation vectorielle

Une fois le plus petit élément à traiter identifié, il nous faut choisir une méthode pour le représenter et le manipuler facilement. En effet, si la phrase, en tant que succession de symboles (mots, caractères, ponctuation, etc.), est déjà une représentation formalisée de la communication (et plus précisément de la langue). Cette représentation est peu satisfaisante pour la tâche qui nous intéresse.
La phrase, succession de mots, n'est pas un objet que l'on peut aisément manipuler ou

14. Même si, comme nous le verrons plus loin dans cette cette section, la notion de segment thématique doit être définie en détail
15. Même si certains cas particuliers de phrases demandent à être désambiguïsés
16. Toutefois, la segmentation en phrases fait l'objet de recherches spécifiques et peut se révéler une tâche complexe si l'on ne veut faire aucune erreur ([Palmer & Hearst, 1997]).

comparer à d'autres au travers d'un modèle ou d'un algorithme. Afin de faciliter nos travaux nous avons décidé d'opter pour une représentation vectorielle du texte. Ce type de représentation se prête beaucoup mieux aux manipulations et aux comparaisons évoquées précédemment. Dans la mesure où nous avons choisi la phrase comme unité atomique, il nous faut représenter chaque phrase par un vecteur [17].
Plusieurs choix s'offraient à nous en matière de représentation vectorielle de la phrase. Cette section s'attache à décrire les principaux choix ayant retenu notre attention et à justifier notre décision d'adopter les vecteurs sémantiques.

3.2.2.1 Le modèle de représentation vectorielle de Salton

Salton, Wong et Yang ([Salton *et al.*, 1975]) ont proposé au milieu des années 70 un modèle de représentation vectorielle du texte qui reste à ce jour toujours très utilisé et souvent désigné sous le nom de modèle de Salton ou de représentation de Salton.
Ce modèle, appelé **Vector Space Model** (ou *term vector model*), est un modèle algébrique utilisé pour représenter un document textuel (mais il peut être utilisé pour représenter des entités textuelles plus petites, comme la phrase dans notre cas) par un vecteur d'identifiants, comme par exemple des termes. C'est un modèle qui a beaucoup été utilisé en recherche d'information, en indexation de documents ou encore pour faire du filtrage d'information. Il a été utilisé pour la première fois dans le cadre du système de recherche d'information SMART ([Salton & Lesk, 1965]).
Son principe est simple, un document est représenté par un vecteur dont chacune des dimensions est un terme [18]. Si ce terme apparaît dans le document, alors la valeur associée à cette dimension est non-nulle. Il existe plusieurs méthodes de calcul de cette valeur, la plus courante restant sans aucun doute la méthode $TF-IDF$ [19]. La valeur d'une dimension est alors la fréquence du terme qui lui correspond au sein du document pondérée par l'inverse de sa fréquence au sein de l'ensemble des documents ayant servi à constituer l'espace vectoriel.
Dans un tel modèle, le vecteur d'un document d aurait cette forme :

$$v_d = [w_{1,d}, w_{2,d}, w_{3,d}, ..., w_{N,d}]^T \tag{3.1}$$

avec $w_{t,d}$:

$$w_{t,d} = tf_t \times log\left(\frac{|D|}{|t \in d|}\right) \tag{3.2}$$

17. Nous parlons bien ici de représentation de la phrase, et non de son traitement.
18. On ne considère en général que les termes dits « utiles » et l'on ignore donc les mots comme certains pronoms ou déterminants par exemple. De plus, ces mots sont lémmatisés pour éviter que les formes fléchies d'un mot ne soient considérées comme des termes différents
19. *Term Frequency - Inverse Document Frequency*

où

- tf_t est la fréquence du terme t au sein du document t.
- $log\left(\frac{|D|}{|t \in d|}\right)$ est l'inverse de la fréquence inter-documentaire de t. $|D|$ étant l'IDF, soit le nombre total de documents ayant servi à construire l'espace vectoriel et $|t \in d|$ le nombre de documents contenant au moins une fois le terme t [20].

La représentation la plus courante du modèle vectoriel de Salton reste une matrice d'occurrences qui met en relation les documents et les termes qu'ils contiennent.

Le modèle vectoriel proposé par Salton, Wong et Yang a toutefois plusieurs inconvénients. Parmi les plus notables :

- Si l'ensemble des documents servant à créer l'espace vectoriel est volumineux et varié, la représentation de chaque document va en pâtir. En effet, de nombreux termes présents impliquent de nombreuses dimensions et donc des vecteurs particulièrement creux pour chaque document individuellement. Dès lors, les comparaisons entre vecteurs, ainsi que les différentes manipulations sur ces vecteurs perdront, respectivement de leur pouvoir discriminant et de leur efficacité.
- Une sensibilité purement lexicale lors de comparaisons. On peut tout à fait imaginer deux documents traitant du même sujet en utilisant un vocabulaire différent. Cette distance entre le champ lexical et le champ sémantique nous intéresse tout particulièrement. En effet, notre tâche s'intéresse plus à la proximité (ou distance) sémantique entre deux portions de texte, qu'à leur proximité (ou distance) lexicale.

Pour être utilisé dans le cadre de la segmentation thématique, le modèle de représentation que nous venons de présenter doit être légèrement adapté. Ainsi, les dimensions de l'espace vectoriel ne correspondent plus aux termes d'un ensemble de documents, mais aux termes du seul document à segmenter. La construction des vecteurs s'effectue selon le même principe que présenté précédemment, à la seule différence que les « documents » sont maintenant des phrases.
Le problème de ce modèle de représentation appliqué à notre tâche de segmentation thématique, est que la réduction des ensembles de référence [21] accentue les problèmes propres à ce modèle que nous avons évoqués plus haut. En effet, appliqué de la sorte ce modèle

20. Il existe d'autres formules de l'IDF, celle présentée n'étant que la plus utilisée. Nous n'avons pas jugé nécessaire de toutes les présenter et les comparer dans la mesure où n'utilisons pas cette représentation.
21. L'ensemble des documents devient un simple document et le document devient une phrase

produit des vecteurs très creux, une phrase pouvant ne contenir qu'un ou deux termes « utiles », mais surtout le problème du caractère purement lexical de cette représentation devient très problématique. Si la probabilité que deux textes traitant du même sujet n'aient aucun mot en commun est faible, ce n'est pas du tout le cas à l'échelle de la phrase. Il est tout à fait envisageable de trouver dans un texte deux phrases traitant exactement du même sujet et n'ayant pourtant aucun terme en commun. Ceci, d'autant plus lorsqu'une règle tacite incite les auteurs à utiliser un maximum de synonymes pour ne pas se répéter [22].

Pour contourner ces difficultés, des extensions de ce modèle ont été proposées. Parmi elles, on peut citer la possibilité d'enrichir les termes du modèle grâce à des ressources lexicales externes telles que des dictionnaires de synonymes ou des réseaux lexicaux comme Wordnet ([Kozima, 1993]). On constate toutefois assez rapidement que si l'enrichissement des termes grâce à un dictionnaire des synonymes ou un réseau lexical permet de surmonter le problème de la synonymie, en contrepartie on voit apparaître un nouveau problème qui est celui de la polysémie. Plus on enrichit le modèle, plus il devient difficile de différencier deux phrases entre elles. L'enrichissement par des termes souvent polysémiques crée des rapprochements entre des phrases sans rapport normalement.

L'exemple simplifié qui suit illustre bien les faiblesses de cette représentation :

Le chat noir a faim et capture une souris. Le sombre félin dévore le rongeur avec appétit.

Cet exemple illustre bien la faiblesse de ce modèle, surtout dans le cadre de notre tâche

	chat	noir	faim	capturer	souris	sombre	félin	rongeur	appétit
Phrase 1	1	1	1	1	1	0	0	0	0
Phrase 2	0	0	0	0	0	1	1	1	1

FIGURE 3.1 – Matrice d'occurrence de l'exemple

de segmentation thématique. Il est évident pour le lecteur que les deux phrases traitent du même sujet. En vérité, leurs énoncés respectifs sont presque identiques sur le plan sémantique. Toutefois si l'on compare ces deux phrases avec une mesure de similarité classique comme le cosinus [23], la mesure considérera les deux phrases comme orthogonales (dans le cas présent le cosinus vaut 0 soit un angle de $\frac{\pi}{2}$).

Si l'on devait « segmenter » cet exemple en ne nous appuyant que sur l'information lexicale, alors on séparerait les deux phrases pour en faire deux segments thématiques différents. A l'évidence, ce serait une erreur.

La représentation de Salton a, bien entendu, fait l'objet de travaux visant à combler ses lacunes, parmi les solutions proposées, une a attiré notre attention : LSA.

22. Ce qui est effectivement le cas en français.
23. Nous nous intéresserons plus en détail à la notion de distance / similarité dans le chapitre suivant.

3.2.2.2 LSA

LSA, pour **Latent Semantic Analysis** (soit Analyse Sémantique Latente) est une technique d'analyse du langage initialement conçue pour l'indexation automatique très utilisée en psychologie ([Deerwester *et al.*, 1990]) et plus précisément en psycholinguistique ([Landauer & Dumais, 1997], [Landauer *et al.*, 1998]). Son objectif est de construire un espace plutôt sémantique que lexical en se basant sur une analyse statistique de l'ensemble des cooccurrences du ou des textes à analyser.
LSA se base sur une propriété des matrices rectangulaires identifiée par Eckart et Young dès 1936 : la décomposition en valeurs singulières. En ce sens, LSA peut être perçue comme une amélioration du modèle de Salton, dans la mesure où cette représentation se trouve être une décomposition en valeurs singulières de la matrice d'occurrences propre au modèle vectoriel de Salton.
La première étape de LSA sera donc de construire une matrice d'occurrences, puis d'effectuer sur celle-ci une décomposition en valeurs singulières. Le résultat de cette décomposition sera le nouvel espace vectoriel que l'on réduira à ses valeurs singulières les plus significatives et dans lequel on « projettera » les documents (ou dans notre cas les phrases). Les dimensions de ce nouvel espace ne correspondant plus aux termes de l'ensemble des documents (ou de l'ensemble des phrases si l'on réduit l'analyse à un unique document), mais à une combinaison de termes. L'objectif est de faire ressortir les liens entre les termes en se basant sur les cooccurrences de ces derniers, des liens qui sont sémantiques et latents [24]. On peut donc employer LSA pour diverses raisons :

- La matrice d'occurrences est trop volumineuse pour que la machine puisse la traiter. La réduction de l'espace à ses valeurs singulières les plus significatives permettant un gain conséquent de place tout en minimisant la perte d'information. La capacité et la puissance des machines modernes ne cessant de croître, c'est probablement la moins « valable » des raisons.
- Le vocabulaire de l'ensemble étudié est très varié (usage de synonymes en grand nombre) et donc la matrice d'occurrences est difficilement analysable. Le nouvel ensemble sera un espace de « concepts » reliant plusieurs termes entre eux, de fait LSA permet d'atténuer les problèmes liés à la synonymie.
- L'espace de termes de l'ensemble étudié est trop « bruité », il contient de nombreux termes anecdotiques ou mal orthographiés par exemples. La réduction à un espace de concept permet de limiter l'effet de ces termes, puisque les valeurs singulières leurs correspondant seront probablement peu élevées et donc éliminées rapidement [25].

24. D'où la dénomination d'Analyse Sémantique Latente.
25. Si les valeurs singulières qui leurs sont associées étaient élevées, les termes n'auraient plus rien d'« anecdotiques ».

LSA étape par étape :
Si l'on considère une matrice X où chaque élément (i, j) décrit les occurrences du terme i (fréquence, TF-IDF, etc.) au sein du document j (ou de la phrase dans notre cadre d'application). X sera alors notre matrice d'occurrence, et correspondra à une représentation de Salton tel que :

$$\begin{array}{c} \\ t_i^T \end{array} \begin{array}{c} \begin{array}{ccc} & d_j & \end{array} \\ \begin{pmatrix} \mathrm{x}_{1,1} & \cdots & \mathrm{x}_{1,n} \\ \vdots & \ddots & \vdots \\ \mathrm{x}_{m,1} & \cdots & \mathrm{x}_{m,n} \end{pmatrix} \end{array} \tag{3.3}$$

Les lignes de cette matrice sont donc des vecteurs donnant l'importance du terme correspondant dans chacun des documents de l'ensemble des textes étudié (ou l'importance du terme correspondant dans chacune des phrases du texte étudié dans notre cadre d'application) :

$$t_i^T = (x_{i,1}, \cdots, x_{i,n}) \tag{3.4}$$

Les colonnes de cette matrice sont des vecteurs représentant chacun un document (ou une phrase si nous travaillons à l'échelle du texte) et donc chaque composante donne l'importance de chacun des termes de l'ensemble étudié pour ce document précis (représentée par une fréquence, un TF-IDF ou toute autre mesure) :

$$d_i = \begin{pmatrix} \mathrm{x}_{1,j} \\ \vdots \\ \mathrm{x}_{m,j} \end{pmatrix} \tag{3.5}$$

Ces vecteurs nous permettent de calculer la corrélation entre termes ou entre documents. Ainsi, le produit scalaire $t_i^T t_p$ donne la corrélation entre le terme i et le terme p. On peut donc considérer le résultat du produit matriciel XX^T comme la matrice contenant l'ensemble des corrélations terme à terme. Cette matrice est symétrique.
On peut procéder de la même manière pour obtenir les corrélations document à document, on considérera alors le produit matriciel $X^T X$.
Effectuer dès lors une décomposition en valeur singulière de la matrice X, nous donne deux matrices orthonormales U et V et une matrice diagonale Σ telles que :

$$X = U\Sigma V^T \tag{3.6}$$

Les produits matriciels permettant d'obtenir les corrélations entre termes ou entre documents s'écrivent alors respectivement :

$$XX^T = (U\Sigma V^T)(U\Sigma V^T)^T = (U\Sigma V^T)(V^{TT}\Sigma^T U^T) = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T \quad (3.7)$$

$$XX^T = (U\Sigma V^T)^T(U\Sigma V^T) = (V^{TT}\Sigma^T U^T)(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T \quad (3.8)$$

Comme $\Sigma\Sigma^T$ et $\Sigma^T\Sigma$ sont diagonales, U est composée des vecteurs propres de XX^T et V des vecteurs propres de X^TX. Les deux produits ont donc les mêmes valeurs propres non-nulles, ces dernières correspondent aux cœfficients diagonaux non-nuls de $\Sigma\Sigma^T$. On obtient donc la décomposition en valeurs singulières :

$$\begin{array}{ccccccccc} & & X & & & U & & \Sigma & & V^T \\ & & (\mathbf{d}_j) & & & & & & & (\hat{\mathbf{d}}_j) \\ & & \downarrow & & & & & & & \downarrow \\ (\mathbf{t}_i^T) \rightarrow & \begin{pmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{pmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \left(\begin{pmatrix} \\ \mathbf{u}_1 \\ \\ \end{pmatrix} \dots \begin{pmatrix} \\ \mathbf{u}_l \\ \\ \end{pmatrix}\right) & \cdot & \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{pmatrix} & \cdot & \begin{pmatrix} (& \mathbf{v}_1 &) \\ & \vdots & \\ (& \mathbf{v}_l &) \end{pmatrix} \end{array} \quad (3.9)$$

Les valeurs $\sigma_1, \dots, \sigma_l$ sont les valeurs singulières de X, de plus les vecteurs $u_1, \dots, u_l$ sont respectivement singuliers à gauche et à droite.
Seule la $i^{ème}$ ligne de U contribue à t_i, on notera donc ce vecteur $\hat{t_i}$. De même, seule la $j^{ème}$ colonne de V^T contribue à d_j, on notera donc ce vecteur $\hat{d_j}$.
La dernière étape de LSA consiste à délimiter l'espace des concepts en ne prenant que les k valeurs singulières les plus élevées. L'espace des concepts est ainsi une approximation de rang k de la matrice d'occurrences.
Les vecteurs termes et documents peuvent donc être maintenant traduits dans cet espace de concept. Le vecteur $\hat{t_i}$ possède k composantes et correspond à l'importance du terme i dans chacun des k concepts. De même, le vecteur $\hat{d_j}$ donne l'intensité des relations entre le document j et chaque concept. Traditionnellement, cette approximation s'écrit :

$$X_k = U_k\Sigma_k V_k^k \quad (3.10)$$

Les avantages que LSA a sur une représentation de Salton classique sont multiples. La réduction de la taille de l'espace de représentation étant le plus évident. Toutefois, la grande force de LSA est de pouvoir, comme nous l'avons évoqué plus haut, « inférer »

des liens sémantiques entre les différents termes de l'ensemble étudié.
Les dimensions du nouvel espace dit de « concepts » ne correspondent plus à des termes, mais à des combinaisons de plusieurs termes de l'espace initial. Ainsi, si nous reprenons l'exemple précédent :

Le chat noir a faim et capture une souris. Le sombre félin dévore le rongeur avec appétit.

Il est impossible de faire le lien entre *chat* et $félin$, entre *noir* et *sombre*, entre $faim$ et

	chat	noir	faim	capturer	souris	sombre	félin	rongeur	appétit
Phrase 1	1	1	1	1	1	0	0	0	0
Phrase 2	0	0	0	0	0	1	1	1	1

FIGURE 3.2 – Matrice d'occurrence de l'exemple (bis)

appétit, entre *souris* et *rongeur* et *a fortiori* de relier les deux phrases. Si l'on suppose que ces deux phrases font partie d'un ensemble plus grand, et que l'on rajoute une phrase :

Le chat noir a faim et capture une souris. Le sombre félin dévore le rongeur avec appétit. Le chat est un félin.

Dés lors, la représentation de Salton devient :
Dans ce nouvel exemple, on voit bien que les phrases 1 et 2 ont des points communs

	chat	noir	faim	capturer	souris	sombre	félin	rongeur	appétit
Phrase 1	1	1	1	1	1	0	0	0	0
Phrase 2	0	0	0	0	0	1	1	1	1
Phrase 3	1	0	0	0	0	0	1	0	0

FIGURE 3.3 – Matrice d'occurrence de l'exemple étendu

avec la phrase 3 (chacune un mot, mais pas le même). Toutefois, on ne constate toujours pas le lien entre les phrases 1 et 2. La grande force de LSA va être de retrouver le lien sémantique qu'il y a entre la phrase 1 et la phrase 2 grâce à la phrase 3. Ce lien est en quelque sorte « inféré par transitivité ».
La faille de LSA réside dans le fait que l'inférence de liens sémantiques entre termes ou entre documents se fait de manière totalement aveugle. Aussi, LSA peut faire apparaître des relations entre des termes n'ayant aucun point commun en faisant une association par transitivité malheureuse, à cause d'un terme particulièrement polysémique. Si nous prenons par exemple ces trois phrases :

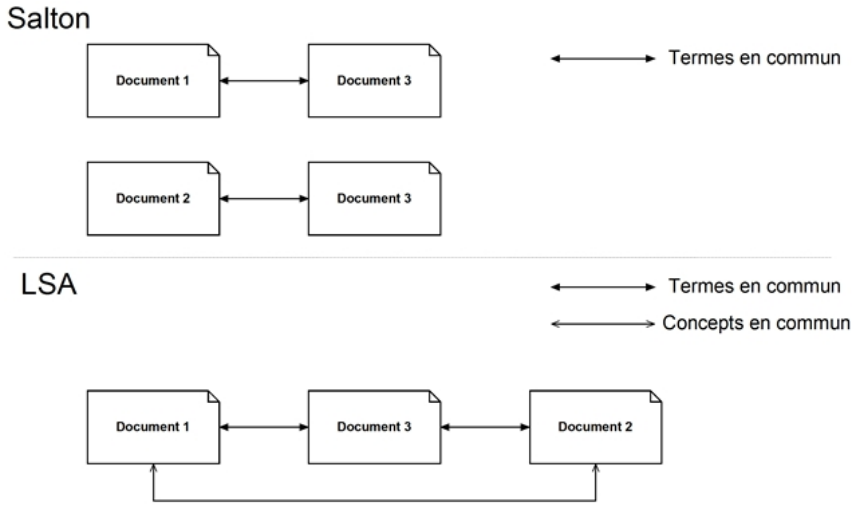

FIGURE 3.4 – La « transitivité » de LSA

- *La glace au chocolat est meilleure que celle à la noix de coco.*
- *Il regarde son reflet dans la glace.*
- *A zéro degrés Celsius l'eau se solidifie pour devenir de la glace.*

Le terme *glace* a une signification différente dans chacune de ces phrases. Il désigne tour à tour un dessert sucré, un objet courant et l'état solide de l'eau. Si les associations que nous obtenons en nous limitant à une représentation de Salton ne sont pas choquantes ($(glace) \rightarrow (chocolat, noix - de - coco)$,$(glace) \rightarrow (reflet)$,$(glace) \rightarrow (eau)$), les association obtenues par « transitivité » grâce à LSA peuvent l'être ($(noix - de - coco) \rightarrow (reflet)$?).
Cette tendance à produire des relations n'ayant pas de sens, ajouté au caractère très lexical de LSA, nous a découragé d'utiliser cette représentation et nous a incité à en trouver une plus détachée du lexique : les vecteurs sémantiques.

3.2.2.3 Vecteur sémantique

Le principe du vecteur sémantique est de projeter la notion linguistique de champ sémantique dans le modèle mathématique d'espace vectoriel. En se basant sur des notions élémentaires, les concepts, identifiés par des linguistes et représentés sous forme de vecteur (dits *vecteurs génératifs*), on peut représenter des segments textuels de toutes tailles (depuis le plus simple item lexical jusqu'au texte complet).
La représentation vectorielle nous dote d'outils mathématiques éprouvés et fiables nous permettant d'effectuer des manipulations sur les portions de texte représentées, formellement bien fondées. Cette représentation, ayant été validée par des linguistes, nous permet

également de formuler des interprétations linguistiques raisonnables. L'hypothèse principale sur laquelle se fonde ce modèle est que l'ensemble des vecteurs génératifs constitue un espace générateur pour l'ensemble des mots de la langue.

Présentation du modèle :
Le principe des vecteurs sémantiques est celui de la linguistique componentionnelle (voir glossaire), c'est à dire que l'on considère que le sens d'un terme peut être décomposé en éléments de sens plus petits. Le modèle que nous présentons est celui proposé et développé par Jacques Chauché depuis le début des années 1990. Il suppose que ces éléments de sens plus petit, désignés par le terme de *concepts*, peuvent être représentés par des vecteurs de $\Re^{+n}$ et qu'ils sont susceptibles de générer l'intégralité des vecteurs sémantiques. Les vecteurs de ces concepts sont appelés *vecteurs génératifs*.
Un vecteur sémantique peut donc être vu comme une combinaison linéaire de *vecteurs génératifs*. Un « sens » serait donc une combinaison linéaire de plusieurs « concepts ». Si l'on considère un ensemble de fini de n concepts $C = \{c_1, c_2, \ldots, c_n\}$, $F = \{V(c_1), V(c_2), \ldots, V(c_n)\}$ l'ensemble des vecteurs correspondant, un item lexical quelconque (mot, proposition, phrase, etc.) l et l'intensité de c_i dans l $\alpha_i \in \Re^i$, nous avons :

$$V(l) = \frac{V}{\|V\|} \text{ où } V = \sum_{i=1}^{n} \alpha_i V(c_i) \tag{3.11}$$

Prenons par exemple comme item lexical le mot <*Football*>. Le vecteur sémantique de cet item lexical peut être construit grâce à une combinaison des concepts SPORT et JEUX comme suit :

$$V(<Football>) = \frac{\alpha_{SPORT} V(SPORT) + \alpha_{JEUX} V(JEUX)}{\|\alpha_{SPORT} V(SPORT) + \alpha_{JEUX} V(JEUX)\|} \tag{3.12}$$

Les vecteurs sémantiques sont donc des vecteurs de $\Re^{+n}$ où n correspond au nombre de concepts identifiés comme élémentaires, et donc au nombre de vecteurs génératifs. Ces vecteurs sont normés à 1 et forment donc une hyper-sphère de rayon 1. n donnant le nombre de concepts, la logique voudrait que plus n est élevé, plus la représentation sera fine. Toutefois augmenter de manière trop importante le nombre de vecteurs génératifs pose deux problèmes majeurs :

- Plus n est élevé, plus la manipulation informatique des vecteurs devient lourde. Même si la puissance des machines ne cesse d'augmenter, il est facile d'imaginer que l'explosion du nombre de concepts, et donc de vecteurs génératifs, peut rapidement poser des problèmes en terme de mémoire et de puissance de calcul.

– Augmenter le nombre de concepts à l'excès reviendrait à avoir un nombre de concepts, et donc un vecteur génératif, supérieur ou égal au nombre de mots et donc perdre tout le bénéfice de cette représentation.

Nous décrirons en détail la génération des vecteurs sémantiques par l'application SYGFRAN dans le chapitre 4.

3.2.2.4 La notion de distance thématique

Comme nous l'avons déjà évoqué, un des avantages d'une représentation vectorielle est la possibilité d'utiliser un certain nombre d'outils mathématiques. Notre tâche consistant à différencier des ensembles textuels parfois très proches il nous faut disposer d'une distance. Nous appellerons « distance thématique » la distance entre deux segments thématiques. Nous évoquerons les différentes distances possibles dans le chapitre 4. Pour l'instant nous nous contenterons de désigner par distance thématique la distance mathématique (quelle qu'elle soit) qui sépare deux vecteurs sémantiques.

3.3 Le segment thématique

Avant même de se poser la question « Comment segmenter thématiquement un texte ?», on doit se demander ce qu'est un segment thématique. Si nous nous référons à la définition du thème que nous avons choisi d'utiliser dans ces travaux, un segment thématique doit être une unité textuelle ne « parlant » que d'un seul et unique sujet.
Toutefois, la notion de « ne parler que d'un seul sujet » est extrêmement vague et subjective. Dans leurs travaux sur la structure du discours [Grosz & Sidner, 2002] reconnaissent une imbrication successive de segments de discours. Dans une tâche automatique comme la nôtre, il nous faut tout de même choisir une échelle standard pour travailler, même si nous ne définissons pas numériquement et explicitement la taille d'un segment thématique « typique ».
Nous nous retrouvons donc face à un problème de granularité. Quelle taille doit faire un segment thématique ?

3.3.1 Le document / le texte dans son intégralité

Dans la mesure où nous souhaitons segmenter un texte en sous-unités textuelles, on peut difficilement faire plus gros en terme de segment thématique que prendre l'intégralité du texte (à part peut être une collection entière de documents traitant d'un même sujet, ou le rayonnage de la bibliothèque, mais cela correspondrait à une tout autre tâche). Toutefois il n'est pas illégitime de considérer un document comme un segment thématique. En

effet, un document traite, la plupart du temps, d'un sujet précis. Le retourner dans son intégralité n'est donc pas une erreur, c'est d'ailleurs ce que font la plupart des moteurs de recherche en réponse à une requête.
Toutefois notre objectif étant de diviser le document en sous parties plus facilement exploitables, on peut se poser la question de savoir si une telle échelle est justifiée. Nous considérerons donc dans ces travaux, qu'un document n'est pas un segment thématique, et que retourner un document non segmenté constituera un échec dans le traitement de la tâche de segmentation thématique.

3.3.2 Le chapitre

Dans le cadre de très gros documents, tels que des livres par exemple, le chapitre est une sous-unité thématique tout à fait acceptable comme segment thématique. En effet, par définition, un chapitre se doit de développer un sujet particulier.
Nous sommes toutefois dans le cas d'une unité textuelle volumineuse. Il est probable que le chapitre ait un fil conducteur unique, mais ce thème sera probablement généraliste et inclura donc plusieurs sous-thèmes plus spécifiques. Par exemple, le chapitre que vous êtes en train de lire traite des notions de détection de changement de thème et d'optimisation de la cohérence thématique. La partie courante traite plus spécifiquement de la notion de segment thématique. Cette notion plus spécifique peut être vue comme un thème à part entière. Le chapitre reste donc trop volumineux pour être un segment thématique et nous allons nous efforcer de définir une taille inférieure pour nos segments thématiques.

3.3.3 La partie / la section

Encore une étape en dessous du chapitre, la partie (ou la section) se rapproche bien plus de la taille souhaitée pour un segment thématique. Toutefois, elle reste une unité textuelle de taille importante, contenant plusieurs paragraphes ou sous-parties. De fait, elle peut très bien traiter de plusieurs thèmes différents (même si ces derniers sont probablement en rapport les uns avec les autres).

3.3.4 Le paragraphe

La plus petite unité textuelle avant la phrase dans la tâche qui nous intéresse. Le paragraphe semble être le segment thématique « idéal ». Pourtant, il est envisageable qu'un paragraphe traite de plus d'un thème, ou qu'un thème soit traité sur plusieurs paragraphes successifs. Aussi, choisir le paragraphe comme segment thématique type pourrait nous amener à commettre des erreurs.

3.3.5 Typographie et segment thématique

Choisir une de ces unités textuelles comme taille de référence pour un segment, c'est affirmer que la structure thématique d'un texte dépend de sa typographie. Il est indéniable que la typographie d'un texte et sa structure thématique sont liées. En tant qu'outils permettant de structurer de l'information textuelle, les indicateurs typographiques servent de jalons tout au long du texte indiquant pauses, ruptures et transition. Mais ces indications ne sont pas absolues. Elles peuvent parfois être mal utilisées (le langage étant le résultat d'un processus humain, il suit rarement les règles à la lettre), voir être détournées de leur fonction initiale volontairement (pour un effet de style, pour réduire la taille d'une portion de texte jugée trop longue, trop lourde, etc.)
Nous avons donc décidé de nous défaire du cadre restrictif qu'impose la mise en page du document, et de considérer le texte comme un bloc uniforme, faisant abstraction de toute information de type typographique. Si ce parti pris nous prive d'une information utile, il nous permet en contre partie de bénéficier d'une plus grande liberté dans la définition d'un segment thématique.

3.3.6 Définition du segment thématique

Nous avons décidé de définir le segment thématique comme étant : *« La plus petite unité textuelle thématiquement cohérente en son sein et thématiquement distincte des unités textuelles précédentes et suivantes. L'unité atomique du segment thématique est la phrase. »*.
Nous avons choisi cette définition car elle nous permet de définir deux critères caractérisant le segment thématique sur lesquels notre modèle s'appuiera. Ces critères que sont la cohérence interne du segment thématique et la distinction de ce dernier vis à vis des segments adjacents ne sont peut être pas les seuls, mais nous avons décidé de nous appuyer sur ces derniers.
On notera que cette définition représente un idéal, qui n'est pas toujours possible d'atteindre. Dans la mesure un retrouver un segment thématique dans un texte d'après cette définition peut être assimilé à l'optimisation de deux critères ont peu imaginer qu'il existe plusieurs solutions possibles.
Tout comme la phrase, le segment thématique peut être représenté par un vecteur sémantique (nous verrons comment plus tard dans la section 3.5.2).

3.4 Quelques notations

Dans la suite de ce chapitre, nous utiliserons un certain nombre de notations propres pour les notions que nous venons d'aborder. Nous les présentons ici :

- Pour simplifier, nous désignerons une phrase de position i dans un ensemble textuel quelconque et le vecteur sémantique qui la représente par un unique symbole p_i.
- Toujours dans un souci de simplification, un segment thématique et son vecteur sémantique représentatif auront le même symbole $V(n, m)$ où n est l'indice de la première phrase du vecteur sémantique dans l'ensemble textuel étudié et m celui de la dernière.
- La distance thématique entre deux vecteurs sémantiques sera notée $D(\alpha, \beta)$ ou α et β sont des vecteurs sémantiques quelconques (de phrase comme de segment).

3.5 Détecter les changements de thème en identifiant des propriétés simples

En langue française, comme dans toutes les langues, la rédaction d'un texte suit un certain nombre de règles, souvent explicites, mais parfois implicites. Ces règles peuvent être très formalisées, comme les règles de grammaire par exemple. Ou beaucoup moins formelles comme les règles qui régissent la structure rhétorique et thématique d'un texte. Aucun manuel de français ne pose explicitement de plan type d'un texte ou encore de structure type du paragraphe. Tout au plus, nous fournit-on des exemples, des textes types, souvent issus d'auteurs célèbres ou des contre-exemples montrant ce qu'il ne faut absolument pas faire. Aussi, les règles régissant la structuration d'un texte sont, pour la plupart, soit définies par défaut, soit définies de manière empirique.
La principale raison à cela est que la langue naturelle est dite « vivante », elle évolue perpétuellement et donc les règles la régissant de même. Si les règles fondamentales régissant la construction de la langue (grammaire, syntaxe, etc.) évoluent peu et donc peuvent être « figées » dans des tables de la loi comme le Bescherelle, il n'en est pas de même pour les règles régissant l'usage de la langue qui lui est en perpétuelle évolution. Nous nous sommes attachés ici à identifier celles de ces propriétés qui pourraient nous aider à déterminer les frontières thématiques au sein d'un texte.

3.5.1 La position des phrases dans le segment thématique

Toutes les phrases n'ont pas la même importance au sein d'un même segment thématique. Il y en a forcément qui sont plus porteuses d'information thématique que d'autres. En partant de ce principe trivial, nous avons cherché quelles phrases au sein d'une unité thématique cohérente, quelle que soit sa granularité, sont les plus porteuses de sens.
Il est rapidement apparu que les première phrases semblent être plus importantes du point de vue thématique que les suivantes. En effet, afin d'être efficace dans sa communication, l'individu a tendance à annoncer le sujet / le thème qu'il souhaite traiter en premier lieu, puis à développer ce thème en l'illustrant d'exemples ou en détaillant son propos. Cette structure nous est enseignée très tôt dans le système éducatif, ce qui fait d'elle un automatisme dans tout texte correctement construit. Notre hypothèse est que cette structure se retrouve à tous les niveaux de granularité du texte, du livre au segment thématique.

3.5.1.1 Exemple concret

Étudions plus en détail un exemple de segment thématique. La portion de texte qui suit est extraite d'un texte du corpus que nous avons utilisé lors de nos travaux. Ce corpus sera présenté plus en détail dans la partie consacrée aux expérimentations. Pour l'instant nous avons juste besoin de savoir que ce segment a été extrait par un expert humain, et donc qu'il a de bonne chances de correspondre à notre hypothèse sur l'information thématique véhiculée par une phrase en fonction de sa position.

Pour ce qui est de la contribution de la Grande-Bretagne au financement de la communauté, la dernière proposition a été faite par le Chancelier SCHMIDT. Elle allait au-delà de ce qui avait été jusqu'alors proposé par nos collègues. J'ai approuvé cette proposition du Chancelier SCHMIDT. Je l'ai même complétée en allant au-delà de ses propres chiffres ou plus exactement en prolongeant la durée de sa proposition. Cette proposition a été rejetée. Elle allait beaucoup plus loin que ce qui avait été envisagé lors du Conseil de Dublin. Elle allait au-delà de ce que, vraisemblablement, les pays en question et leur gouvernement étaient disposés à consentir comme sacrifice sur-le-plan budgétaire. C'était donc une contribution importante que nous proposions d'apporter ensemble à la solution de ce problème. Cette proposition a été rejetée. La réunion de ce Conseil européen a un certain effet révélateur. Il est apparu que les demandes britanniques, en-raison de leur importance, de leur ampleur, en-raison de leur durée, ne pouvaient pas recevoir de solution dans-le-cadre des règlements communautaires existants. Nous sommes allés jusqu'à la limite de ce qu'on pouvait attendre ou demander à ces réglementations communautaires existantes. Nous avons vu progressivement autour de la table les principales délégations partager ce sentiment.

Cette portion de texte a été considérée comme un segment thématique par l'expert humain qui l'a extraite. Si nous la regardons de plus près, nous constatons qu'elle peut être divisée en trois parties bien distinctes :
Tout d'abord la première phrase.

> *Pour ce qui est de la contribution de la Grande-Bretagne au financement de la communauté, la dernière proposition a été faite par le Chancelier SCHMIDT.*

La première phrase porte visiblement l'information essentielle que l'auteur veut véhiculer. Tout y est, on sait que l'on va parler de la Grande-Bretagne, et plus précisément de sa contribution financière et on sait q'une proposition a été faite sur ce sujet par le Chancelier SCHMIDT.

> *Elle allait au-delà de ce qui avait été jusqu'alors proposé par nos collègues. J'ai approuvé cette proposition du Chancelier SCHMIDT. Je l'ai même complétée en allant au-delà de ses propres chiffres ou plus exactement en prolongeant la durée de sa proposition. Cette proposition a été rejetée.*

Les quatre phrases suivantes portent, elles aussi, beaucoup d'informations, mais l'information commence déjà à se diluer. On y apprend le lien que l'auteur a avec le thème abordé (il a approuvé et complété la proposition), ainsi qu'une information importante, la proposition a été rejetée.

> *Elle allait beaucoup plus loin que ce qui avait été envisagé lors du Conseil de Dublin. Elle allait au-delà de ce que, vraisemblablement, les pays en question et leur gouvernement étaient disposés à consentir comme sacrifice sur-le-plan budgétaire. C'était donc une contribution importante que nous proposions d'apporter ensemble à la solution de ce problème. Cette proposition a été rejetée. La réunion de ce Conseil européen a un certain effet révélateur. Il est apparu que les demandes britanniques, en-raison de leur importance, de leur ampleur, en-raison de leur durée, ne pouvaient pas recevoir de solution dans-le-cadre des règlements communautaires existants. Nous sommes allés jusqu'à la limite de ce qu'on pouvait attendre ou demander à ces réglementations communautaires existantes. Nous avons vu progressivement autour de la table les principales délégations partager ce sentiment.*

Les dernières phrases sont clairement des phrases qui contribuent à la compréhension, mais qui ne sont pas indispensables. On y trouve un développement de l'information, avec, notamment, la justification du rejet de la proposition et quelques éclaircissements sur la nature de la proposition.
Si indubitablement la première phrase est indispensable à la compréhension de l'informa-

tion véhiculée par ce segment thématique, et que les quatre suivantes ajoutent grandement à la compréhension, la fin du segment semble, quant à elle, beaucoup moins importante. Cette structure peut être représentée de la manière suivante (figure 3.5) :

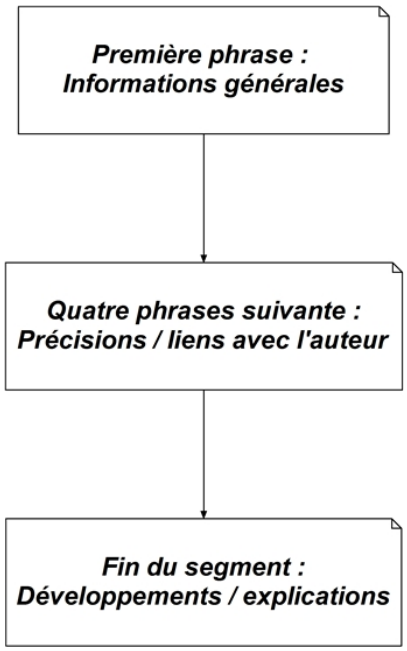

FIGURE 3.5 – Structure de l'exemple

3.5.1.2 Hypothèse sur l'organisation d'un segment thématique

L'exemple précédent fait ressortir la possible existence d'une structure « type » du segment thématique. Cette structure supposerait 3 étapes dans l'organisation d'un segment thématique :

- Les toutes premières phrases (voir juste la première) qui vont exposer le thème du segment. Cette première étape va véritablement définir « ce dont parle » le segment et se suffit généralement à elle-même. Elle apporte l'essentiel de l'information nécessaire à la compréhension du thème, mais reste générale.
- Les phrases suivantes qui vont préciser le thème, l'affiner. La deuxième étape va combler les lacunes laissées par l'information très généraliste fournie par les premières phrases. En spécifiant le thème du segment, cette étape va permettre de différencier plus facilement deux thèmes du même texte, alors que ceux-ci ont de bonne chances d'être proches.
- Les dernières phrases correspondent, en général, à des exemples, des précisions ou même parfois des digressions. Cette étape apporte peu à la compréhension du seg-

ment thématique, mais permet de conclure ce dernier sans brutalité. Souvent, cette étape sert à amorcer le segment thématique suivant, mais ce n'est pas toujours le cas.

La figure 3.6 illustre la structure que nous venons d'évoquer.
On peut rapprocher cette structure des théories des linguistes et précurseurs de la prag-

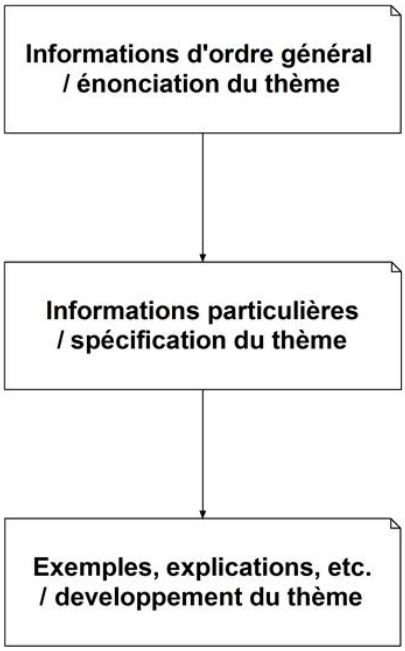

FIGURE 3.6 – Organisation d'un segment thématique

matique, Bühler et Jakobson.
Bühler en 1933 propose dans son modèle « Organon » de séparer le processus de communication en trois fonctions qui correspondent aux trois ingrédients essentiels de la communication :

- La fonction *de représentation* (appelée fonction *référentielle* par Jakobson) qui est corrélée avec le « monde » dans les travaux de Bühler et qui correspondrait dans la structure que nous avons évoquée à la première étape. En effet, la « fonction » de la première étape est de poser le sujet du segment thématique, de « représenter » le thème.
- La fonction *d'expression* (appelée fonction *expressive* par Jakobson) qui est corrélée avec le locuteur dans les travaux de Bühler et qui correspondrait à la deuxième étape de la structure du segment thématique. En précisant le thème abordé, l'auteur ou le locuteur exprime très souvent soit un avis personnel soit son lien avec le thème.
- La fonction *d'appel* (appelée fonction *conative* par Jakobson) qui est corrélée avec le destinataire dans les travaux de Bühler et qui correspondrait à la troisième et

dernière étape de la structure du segment thématique. La fin du segment, au travers d'exemples et de digressions, « ouvre » l'information véhiculée par le segment thématique à son destinataire.

Les phrases d'un segment thématique ont donc une « importance thématique » différente en fonction de leur localisation au sein du segment. Les premières phrases étant les plus importantes, puis leur importance allant en déclinant au fur et à mesure que l'on avance dans le segment. On notera que ce phénomène a été observé dans d'autres travaux sur le domaine tel que ceux de [Lin & Hovy, 1997] ou plus récemment de [Lelu *et al.*, 2006].
Il nous fallait modéliser ce concept de l'importance thématique d'une phrase en fonction de sa position. Il nous fallait définir une « forme » du segment thématique et une manière de le représenter.

3.5.2 La « forme » du segment thématique

En nous basant sur l'hypothèse que nous venons de présenter sur l'importance thématique des phrases au sein d'un segment en fonction de leur position dans ce dernier et en nous appuyant sur notre modèle de représentation vectorielle, nous avons choisi de représenter un segment thématique par un vecteur sémantique qui serait le barycentre des vecteurs des phrases qui le composent.
Ce choix présente l'avantage d'être dans la continuité directe du modèle de vecteur sémantique de phrase que nous avons choisi. Ce modèle fonctionnant par combinaisons linéaires successives, le vecteur sémantique du segment thématique n'est qu'une combinaison linéaire de plus, devenant ainsi une étape supplémentaire dans la représentation du texte.
De plus, le barycentre nous permet d'attribuer un cœfficient différent à chacune des phrases du segment thématique, et donc de représenter mathématiquement l'importance thématique d'une phrase relativement aux autres phrases du segment.
Ainsi, si nous considérons un segment thématique quelconque composé de n phrases p_i, nous aurons $V(0, n-1)$:

$$V(0, n-1) = \frac{\sum_{i=0}^{n-1} \alpha_i p_i}{\sum_{i=0}^{n-1} \alpha_{p_i}} \tag{3.13}$$

Avec α_i le cœfficient représentant l'importance attribuée à la phrase p_i.
Il est donc maintenant aisé d'appliquer une « forme » particulière au segment en jouant sur les paramètres α_i. Dans la mesure où les premières phrases doivent être privilégiées sur les suivantes, nous avons choisi de faire varier la valeur du paramètre α_i en fonction de i selon une fonction mathématique. Plusieurs possibilités s'offraient à nous, nous en

avons exploré trois.

3.5.2.1 La forme affine

La première et la plus simple des fonction que nous avons testée(en dehors peut être de « pas de fonction ») est sans conteste une fonction affine. Cette fonction suppose que les phrases « perdent » de leur importance au sein du segment thématique de manière régulière. La figure 3.7 illustre cette perte graduelle de la valeur thématique de la phrase en fonction de sa position dans le segment.

Si l'on considère un segment de taille n, alors le cœfficient α_i de la phrase p_i sera :

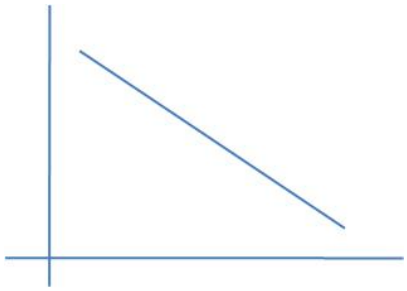

FIGURE 3.7 – Fonction de forme linéaire

$$\alpha_i = n - i \tag{3.14}$$

Avec i variant de 0 à $n - 1$.

3.5.2.2 La forme exponentielle

La fonction de forme exponentielle favorise grandement la première phrase par rapport aux suivantes. Une telle forme suppose que seule la première phrase est porteuse d'information. La figure 3.8 représente ce que serait une telle forme.

La première formule envisagée pour modéliser cette fonction fut pour un segment de

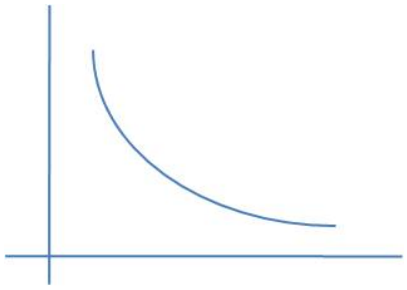

FIGURE 3.8 – Fonction de forme exponentielle

taille n et un cœfficient α_i de la phrase p_i :

α_i=e^{n-i}(3.15)Avec i variant de 0 à $n-1$.
Toutefois, cela donne une décroissance beaucoup trop brutale, réduisant le segment thématique à sa seule première phrase. Nous avons donc opté pour une fonction qui n'a d'exponentielle que le nom, mais qui respecte la « forme » voulue pour le segment thématique. Dans les mêmes conditions que précédemment nous avons :

$$\alpha_i = \sum_{j=1}^{i+1} j \tag{3.16}$$

3.5.2.3 La forme « sinusoïdale »

La forme sinusoïdale suppose que le segment thématique se décompose en trois parties. Les toutes premières phrases qui auraient une grande valeur thématique, les phrases du milieu du segment dont l'importance thématique déclinerait de manière quasi linéaire et les dernières phrases qui elles seraient presque négligeables. La figure 3.9 illustre cette « forme » thématique.

Modéliser une telle forme sur un ensemble de départ borné et de taille variable[26] n'est

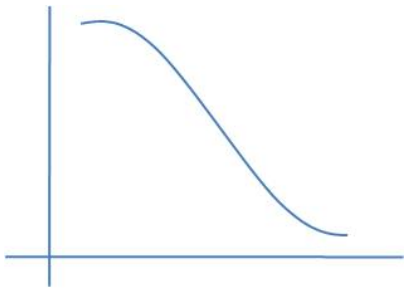

FIGURE 3.9 – Fonction de forme sinusoïdale

pas chose aisée. Après plusieurs tentatives infructueuses ou peu satisfaisantes, nous avons finalement opté pour la formule suivante :

$$\alpha_i = n.\tanh(\frac{n}{2i}) \tag{3.17}$$

Toujours en considérant un segment de taille n, un cœfficient α_{p_i} et une phrase p_i, avec i variant de 0 à $n-1$.
Cette formule s'est avérée la plus proche de la « forme » thématique désirée. Ainsi pour un segment de taille $n = 20$ nous obtenons la forme présentée dans la figure 3.5.2.3.

26. L'ensemble de départ de chacune des fonctions de forme devant logiquement être $[0, n-1]$, où n est le nombre de phrases du segment.

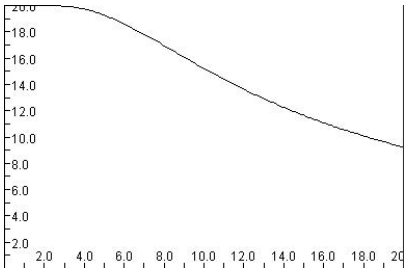

FIGURE 3.10 – Forme « sinusoïdale » pour un segment de 20 phrases

3.5.2.4 Commentaire sur les formes proposées

D'autres formes de segment thématique étaient envisageables, mais les trois que nous avons présentées nous semblaient être les plus pertinentes. Des trois fonction de forme, la forme sinusoïdale est celle qui avait le plus nos faveurs. En effet, des trois c'est celle qui se rapproche le plus de notre hypothèse sur la manière dont un segment thématique s'organise. Nous verrons dans la partie réservée aux expérimentations que parfois les modèles les plus simples sont les meilleurs (section 5.5.2.1).

3.5.3 La notion de zone de transition

Si nous tenons compte des observations précédentes sur la « forme thématique » supposée d'un segment thématique et du fait que le langage naturel est un phénomène continu (même si nous le discrétisons pour les besoins de l'analyse), alors il paraît illusoire de penser que les frontières thématiques d'un texte soient abruptes. Nous avons fait le choix de la phrase comme unité atomique pour nos segments thématiques. Aussi nos frontières seront identifiées par des phrases, mais peut on toujours affirmer qu'une phrase est « plus » une frontière que la phrase qui la précède ou qui la suit. C'est parfois le cas, lors de transitions nettes et radicales au sein du texte. Mais, lorsque plusieurs thèmes s'enchaînent selon un cheminement logique, les frontières deviennent en général plus floues.
Sachant cela, nous avons décidé d'introduire la notion de zone de transition.
Définition : Une zone de transition entre deux segments thématiques est un ensemble de phrases à cheval sur les deux segments concluant le thème du premier segment tout en introduisant le thème du second.
Idéalement, une zone de transition devrait se composer de deux phrases :

- La dernière phrase du premier segment. Celle-ci doit conclure le segment et éventuellement amorcer l'introduction du segment suivant.
- La première phrase du second segment. Celle-ci doit exposer le thème du segment tout en rappelant éventuellement le thème du segment précédent.

Nous nous sommes donc attachés à rechercher les propriétés de ces deux types de phrases au sein de notre modèle de vecteurs sémantiques.

3.5.4 Propriété d'une amorce de segment thématique

La première phrase d'un segment thématique se doit d'introduire le thème du segment. Selon notre représentation, c'est la phrase la plus importante thématiquement du segment. La première phrase sera donc celle qui maximisera la distance thématique entre le segment qu'elle introduit et le segment précédent.
Bien entendu, nous ne connaissons pas les segments thématiques puisque nous les recherchons, vérifier cette propriété n'est donc pas possible *a priori*. Il nous faut donc supposer l'existence d'un segment thématique commençant par la phrase dont nous voulons tester les propriétés, ainsi que l'existence d'un segment thématique qui la précède.
Notre algorithme visant à identifier les première phrases de segments thématiques procède donc en utilisant une fenêtre. Nous faisons glisser cette fenêtre le long du texte, et considérons chaque moitié de la fenêtre comme un segment thématique. En calculant la distance entre ces deux segments supposés nous obtenons un score qui nous permet d'estimer dans quelle mesure la phrase commençant le deuxième segment est la première phrase d'un segment thématique.
Nous appelons ce score : *score de transition* et nous le notons T_i :

$$T_i = D(V(0, i-1), V(i, n-1)) \tag{3.18}$$

où n est la taille de la fenêtre.
Ce score sera conservé dans un tableau pour pouvoir retrouver les zones de transition (figure 3.11).

3.5.5 Propriétés d'une fin de segment thématique

La dernière phrase d'un segment thématique se doit de conclure le thème courant et éventuellement d'introduire le suivant. C'est un comportement idéal qui dans la réalité ne se produit pas toujours. Contrairement aux débuts de segments thématiques qui ne s'écartent que très rarement de leur configuration idéale, les fins de segments thématiques sont moins définies. Souvent, la dernière phrase d'un segment thématique n'est que la dernière phrase d'un exemple assez éloigné du thème du segment ou encore une phrase d'usage qui fait plus office de « ponctuation » thématique que de vraie transition.
Toutefois, que la phrase respecte le modèle que nous avons dessiné ou qu'elle s'éloigne de

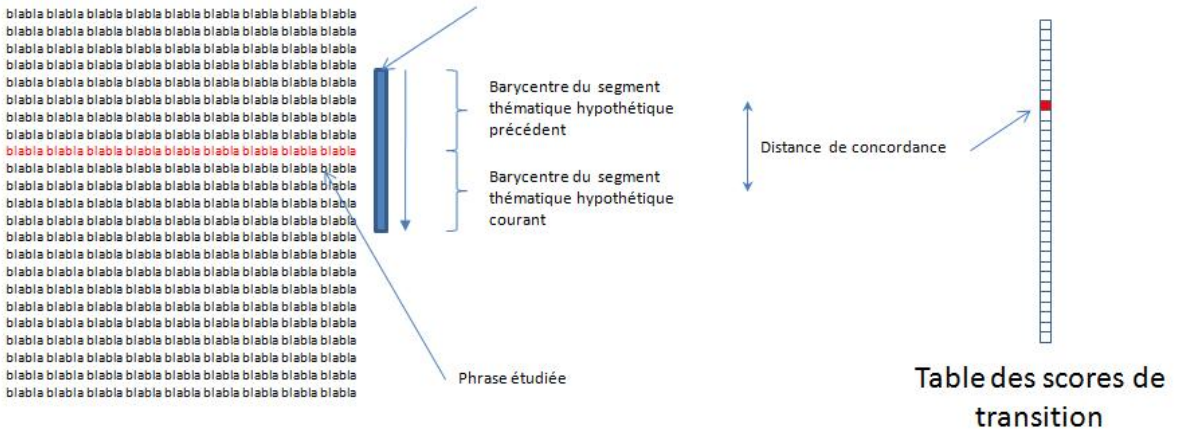

FIGURE 3.11 – Le score de transition d'une phrase

sa fonction, elle conserve tout de même une propriété essentielle : elle est aussi éloignée du segment qu'elle conclut que de celui qui lui succède [27].

Toujours en utilisant notre fenêtre, nous calculons la distance thématique entre la phrase centrale de la fenêtre et chacun des segments potentiels de la fenêtre (à savoir l'ensemble des phrases qui la précèdent et l'ensemble des phrases qui lui succèdent). Plus la valeur absolue de la différence entre ces deux distances est élevée, moins il y a de chances que la phrase étudiée soit une fin de segment thématique.

En effet, si la valeur absolue de cette différence est élevée, cela signifie que la phrase étudiée est beaucoup plus proche d'un segment que d'un autre. Dès lors, elle ne vérifie pas la propriété d'équidistance que nous avons énoncée plus haut. Nous identifions cette propriété par un *score de rupture*, noté R_i, que nous calculons ainsi :

$$R_i = 1 - |D(V(0, i-1), p_i) - D(p_i, V(i+1, n))| \tag{3.19}$$

avec $D(V(0, i-1), p_i)$ la distance thématique entre la phrase centrale de la fenêtre et le segment précédent supposé et $D(p_i, V(i+1, n))$ la distance thématique entre la phrase centrale de la fenêtre et le segment suivant supposé [28]. Le score de rupture a été construit de telle manière qu'il soit le plus élevé possible lorsque la phrase étudiée a le plus de chance d'être une fin de segment.

Comme pour le score de transition le score de rupture est stocké dans un tableau (figure 3.12).

27. Que ce soit en étant proche des deux dans le cas d'une transition « propre » ou loin des deux dans le cas d'une divergence par rapport à notre modèle idéal.

28. On notera que contrairement au score de transition le segment précédent supposé n'inclut pas la phrase étudiée.

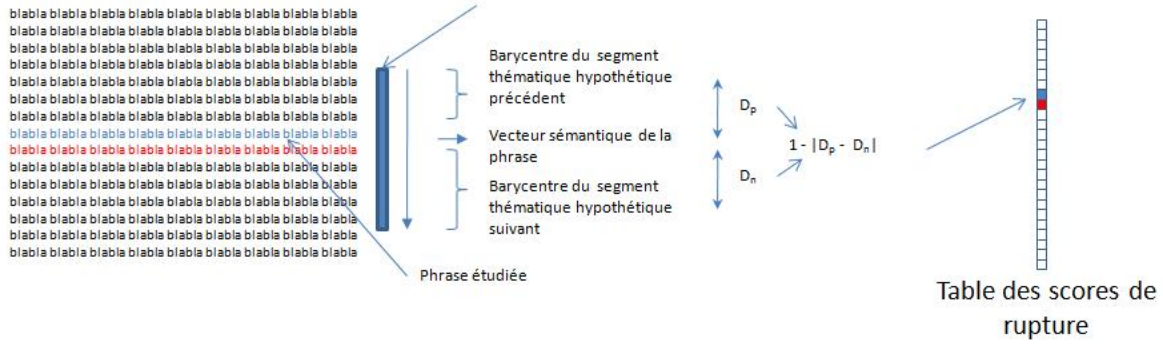

FIGURE 3.12 – Le score de rupture d'une phrase

3.5.6 Identifier les frontières thématiques

La langue est un phénomène continu, et nous avons nous même postulé qu'il n'existait pas de frontières thématiques abruptes, mais des zones de transition. Toutefois, notre tâche nous oblige à désigner une phrase comme frontière entre deux segments thématiques, tout comme elle nous a obligé à discrétiser le phénomène linguistique en considérant la phrase comme un élément atomique.

3.5.6.1 Localiser les zones de transition possibles

Si nous suivons la logique de notre modèle, il nous faut d'abord localiser les zones de transition. Pour ce faire nous utilisons les scores de transition que nous avons calculés pour chacune des phrases du texte. Nous considérerons que les phrases successives ayant un score de transition supérieur à un seuil donné forment une zone de transition. Nous ignorerons les phrases isolées dont le score de transition est supérieur à ce seuil dans la mesure où une zone de transition comporte au moins deux phrases.

Le choix du seuil sera discuté dans un autre chapitre, mais il parait raisonnable d'envisager

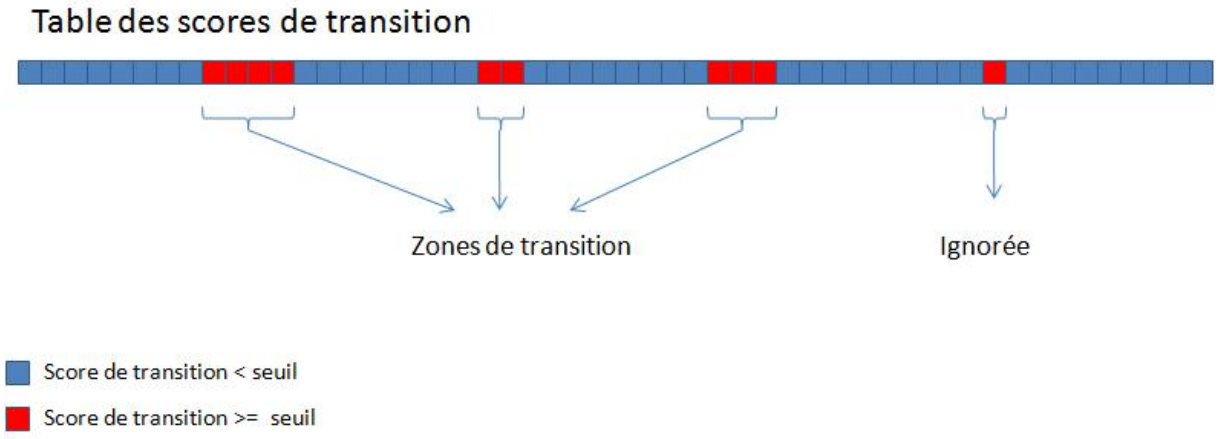

FIGURE 3.13 – Localisation des zones de transition

un seuil de $0,5$ dans le cadre d'une distance thématique qui serait normalisée.

3.5.6.2 Choisir la phrase frontière dans la zone de transition

Une fois les zones de transition localisées, il faut désigner dans chacune d'elles la phrase qui sera la « frontière thématique » [29]. Ce choix se fait grâce au score de rupture.
En combinant le score de transition de chaque phrase d'une zone de transition avec le score de rupture de la phrase qui la précède, nous obtenons un troisième score que nous appelons *score final* et que nous notons F_i. F_i se calcule comme suit :

$$F_i = T_i \times R_{i-1} \tag{3.20}$$

La phrase désignée comme frontière thématique sera la phrase dans la zone de transition dont le score final F_i sera le plus élevé.

3.6 Rechercher la cohérence thématique dans un texte en utilisant le clustering

Si nous considérons la définition d'un segment thématique que nous avons donnée au début de ce chapitre, la section précédente, en s'intéressant très spécifiquement aux propriétés des frontières, s'est attachée à la deuxième des propriétés du segment thématique, sa distinction thématique vis à vis des segments qui l'entourent. Dans cette section nous allons nous intéresser à l'autre propriété énoncée dans la définition, la cohérence des phrases qui le composent.
Il s'agit donc de regrouper les phrases selon deux critères :

- Les phrases doivent être thématiquement proches.
- Les phrases doivent se suivre dans l'ordre du texte.

Le deuxième critère étant trivial dans son traitement, nous avons décidé de voir quelle méthode serait la plus adaptée pour atteindre le premier critère. La solution qui nous est apparue la plus cohérente nous est venue du domaine de la fouille de données : le clustering.

Définition : Le clustering est une technique de fouille de données consistant à regrouper des éléments de données dans des catégories qui n'auront été ni définies ni apprises en amont de la tâche.

29. Voir glossaire.

Cela signifie que les algorithmes de clustering regroupent les éléments proches et séparent les éléments différents dans un ensemble défini d'éléments, et cela sans apprentissage supervisé. Notre tâche consistant à regrouper les phrases thématiquement proches entre elles sans avoir d'information autres que les autres phrases du texte, l'approche par clustering semble appropriée.
Une fois les phrases regroupées en clusters, il ne reste plus qu'à regarder celles qui appartiennent à la même classe tout en étant successives pour former des segments thématiques.

3.6.1 La problématique de la détermination automatique du nombre de thèmes

La principale difficulté que pose l'approche par clustering est la détermination *a priori* du nombre de classes, donc du nombre de thèmes. En effet, si les algorithmes de clustering ne demandent pas d'information sur les clusters, il faut en général paramétrer le nombre de clusters (de thèmes dans notre cas) *a priori*[30].
Le problème qui s'est donc posé à nous était de savoir combien de thèmes un texte possède avant même de l'avoir segmenté thématiquement. Nous avons envisagé deux approches que nous décrivons ici.

3.6.1.1 L'approche « naïve »

La solution la plus simple à ce problème serait de naïvement dire que nous n'avons pas vraiment besoin de savoir combien de thèmes sont abordés par un texte. En effet, notre segmentation étant linéaire, seul deux clusters sont nécessaires pour que l'approche par clustering fonctionne. Ainsi, les segments thématiques seraient les phrases successives appartenant alternativement à chacun des deux clusters.
Si cette approche à le mérite d'être extrêmement simple, son principal défaut et qu'elle risque de « rater » des frontières thématiques en regroupant des thèmes proches, mais différents du fait du nombre réduit de clusters.

3.6.1.2 L'approche statistique

L'autre solution est d'approximer le nombre de thèmes du texte en se basant sur les informations qu'il nous fournit. La représentation en vecteurs sémantiques que nous faisons du texte nous permet de disposer d'une information utile pour résoudre notre problème : les concepts.
En effet, chaque valeur des vecteurs sémantiques correspond à un concept identifié par les

30. L'algorithme XMean développé par [Pelleg & Moore, 2000] n'a pas ce problème, nous le traitons à part dans cette section

linguistes qui ont conçu le thésaurus (Larousse [Larousse, 1992b] dans notre cas). Ce ne sont ni des mots, ni des artefacts obtenus à la suite d'une analyse statistique. Ce sont des concepts concrets et même s'il est entendu qu'un concept n'est pas un thème, le concept du thésaurus et le thème du texte sont fortement liés.
En partant du principe qu'un thème se compose de plusieurs concepts, nous avons postulé que l'un des concepts composant chaque thème doit être plus représenté que les autres. Ces concepts que l'ont appellera « directeurs » nous serviront à estimer le nombre de thèmes. Si ces concepts directeurs sont les principaux concepts des thèmes, ils doivent être plus présents dans le texte que les autres.
L'estimation se fera donc en sommant toutes les valeurs de chacun des concepts sur la totalité du texte. Une fois ces valeurs, que nous appellerons valeurs d'activation, obtenues, nous calculons des maxima locaux (en passant par la dérivée) et considérons les maxima locaux au dessus d'un certain seuil comme étant des concepts directeurs. Le choix de se seuil étant problématique, nous avons choisi la somme de la moyenne des valeurs d'activation avec un nombre λ de fois l'écart type. Ce seuil est appelé seuil de saillance, on le notera s.

$$s = M + \lambda\sigma \tag{3.21}$$

avec M la moyenne des valeurs d'activation totale de chaque composante et σ leur écart type.

Le problème de cette approche est triple.

- D'abord, rien n'empêche un thème d'avoir plus d'un concept directeur, en fait c'est même souvent le cas. Ainsi, notre approche aura tendance à trouver trop de thèmes. On pourrait penser que ce travers entraînerait une sur-segmentation [31], mais en réalité les algorithmes de clustering sont ainsi conçus qu'ils peuvent produire des résultats avec des clusters vides. Les initialiser avec un trop grand nombre de « thèmes » n'a donc pas nécessairement beaucoup d'impact sur le résultat.
- La détermination du λ et donc du seuil est elle aussi problématique. Quelle valeur choisir ? Lors de nos expériences nous avons testé plusieurs valeurs possibles et nous nous sommes arrêtés sur $\lambda = 3$ qui donne un nombre de thèmes inférieur à 20, $\lambda = 4$ donnait des valeurs trop petites (en général 1, parfois 2). Ce tâtonnement empirique n'est toutefois pas satisfaisant.
- Cette approche suppose que la répartition des concepts au sein d'un thème est modélisée par une loi normale. Hors rien n'est moins sûr.

31. Un trop grand nombre de thèmes séparant des phrases qui devraient être ensembles.

Malgré ces problèmes, cette solution se révèle plus satisfaisante que l'approche naïve.

3.6.2 Le clustering strict comme approche de segmentation thématique non supervisée

La première option que nous avons étudiée fut le clustering strict, c'est à dire qui ne considère pas qu'un élément de l'ensemble des données puisse appartenir à plus d'un cluster. Pour tester l'efficacité du clustering strict nous avons employé un algorithme connu et très utilisé en fouille de données : l'algorithme K-Means (ou K-Moyennes).
L'algorithme K-Means partitionne N éléments d'un ensemble de données E en k sous-ensembles (clusters) S_j contenant N_j éléments en minimisant le critère de la somme des carrés J suivant :

$$j = \sum_{j=1}^{k} \sum_{n \in S_j} |x_n - \mu_j|^2 \tag{3.22}$$

avec x_n le vecteur représentant le $n$ième élément et μ_j le centroïde de l'ensemble des éléments composant S_j.
Le principal défaut de cet algorithme et de n'atteindre en général qu'un minimum local, l'initialisation des clusters se faisant de manière aléatoire. Il est donc nécessaire de faire plusieurs itérations de l'algorithme en utilisant le résultat de l'itération précédente comme initialisation pour l'itération suivante.
Un autre défaut, moins technique celui là, est la vision qu'un clustering strict donne du texte et de la tâche de segmentation thématique. En cloisonnant les phrases dans des clusters fermés on suppose que les phrases sont indépendantes les unes des autres. Hors, il parait évident que les phrases d'un texte ne sont pas indépendantes les unes des autres par construction.

3.6.3 X-Mean : une amélioration de K-Mean

Face au problème de la détermination du nombre de clusters que pose l'algorithme K-Mean, [Pelleg & Moore, 2000] ont proposé une amélioration de ce dernier qui détermine automatiquement le nombre de clusters : X-Mean.
Plutôt que de demander à l'utilisateur le nombre de clusters (k), l'algorithme X-Mean lui demande une fourchette dans laquelle il est raisonnable de penser que k se situe. Ensuite, l'algorithme procède de manière assez simple :

- Initialisation de k à la valeur minimum de la fourchette.
- Tant que k reste dans la fourchette

  - On fait converger K-Mean avec le k courant.
  - On estime la valeur du résultat du clustering grâce à une mesure (BIC scoring).
  - On détermine où le prochain centroïde doit apparaître.
  - On incrémente k

- On renvoie le résultat ayant eu la meilleure estimation.

L'algorithme X-Mean conserve les autres défauts de K-Mean, mais propose une solution pour remédier au défaut général des algorithmes de clustering, à savoir la détermination empirique du nombre de clusters.

3.6.4 Le clustering flou comme approche de segmentation thématique non supervisée

La tâche de clustering flou est globalement la même que celle de clustering strict à une différence près. Dans un clustering flou on envisage qu'un élément de l'ensemble de données puisse appartenir à plusieurs clusters. Au lieu de ranger un élément de l'ensemble de données dans un cluster précis, un algorithme de clustering flou attribue à chaque élément de l'ensemble de données une probabilité que ce dernier appartienne à chacun des clusters.
Regrouper en segments thématiques les phrases sur la base des probabilités données par un algorithme de clustering flou se fait de la même manière que pour un algorithme strict à quelques différences près. Comme nous disposons de probabilités d'appartenance à un thème plutôt que d'une appartenance stricte pour chaque phrase, il nous est possible d'identifier certaines formes particulières :

- Si un petit groupe de phrases (une ou deux) appartient de manière peu significative à un thème (la probabilité est supérieure aux autres, mais pas de beaucoup) et qu'il est au milieu d'un groupe de phrases appartenant à un autre thème, alors il y a des chances qu'il s'agisse là d'une digression ou d'une erreur de l'algorithme. On peut donc assimiler ce petit groupe au groupe qui l'englobe et ne former plus qu'un seul groupe au lieu de trois.
- Si on observe une alternance rapide de deux thèmes (sur des segments d'une ou deux phrases) et que leurs probabilités respectives pour chaque phrase ne sont pas très éloignées, alors il est probable que les deux thèmes ne fassent en réalité qu'un.

On notera que la somme des probabilités d'appartenance à chacun des clusters pour un élément de l'ensemble de données n'est pas égale à 1 dans la plupart de ces algorithmes de clustering flou.

L'algorithme de clustering flou que nous avons employé est un algorithme nommé Espérance-Maximisation (que l'on nomme algorithme EM par souci de concision). C'est sûrement le plus populaire et le plus utilisé des algorithmes de clustering flou car il a le double avantage d'être efficace et relativement simple dans son principe.
L'algorithme EM alterne de manière itérative des phases d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

L'algorithme EM

Si l'on considère un échantillon $X = (x_1, \ldots, x_n)$ d'individus suivant une loi $f(x_i; \theta)$ paramétrée par θ, on cherche à déterminer le paramètre θ maximisant la log-vraisemblance donnée par :

$$L(X; \theta) = \sum_{i=1}^{n} log\ f(x_i; \theta) \tag{3.23}$$

Pour déterminer le paramètre θ, il est nécessaire de s'appuyer sur des données complétées par un vecteur inconnu $Z = (z_{1,\ldots,z_n})$. En notant $f(z_i|x_i; \theta)$ la probabilité de z_i sachant x_i et le paramètre θ, on peut définir la log-vraisemblance complétée comme la quantité :

$$L((X, Z); \theta) = \sum_{i=1}^{n} (logf(z_i|x_i; \theta) + logf(x_i; \theta)) \tag{3.24}$$

A partir de quoi, on peut en déduire :

$$L(X; \theta) = L((X, Z); \theta) + \sum_{i=1}^{n} log\ f(z_i|x_i; \theta) \tag{3.25}$$

L'algorithme EM est une procédure itérative basée sur l'espérance des données complétées conditionnellement au paramètre courant. En notant $\theta^{(c)}$ ce paramètre, on peut écrire :

$$E[L(X; \theta)|\theta^{(c)}] = E[L((X, Z); \theta)|\theta^{(c)}] - E[\sum_{i=1}^{n} (logf(z_i|x_i; \theta)|\theta^{(c)}] \tag{3.26}$$

ou encore :

$$L(X;\theta) = Q(\theta;\theta^{(c)}) - H(\theta;\theta^{(c)}) \tag{3.27}$$

avec :

$$Q(\theta;\theta^{(c)}) = E[L((X,Z);\theta)|\theta^{(c)}] \tag{3.28}$$

et

$$H(\theta;\theta^{(c)}) = E[\sum_{i=1}^{n}(logf(z_i|x_i;\theta)|\theta^{(c)}] \tag{3.29}$$

On peut montrer que la suite définie par :

$$\theta^{(c+1)} = argmax_\theta(Q(\theta;\theta^{(c)})) \tag{3.30}$$

fait tendre $L(X;\theta^{(c+1)})$ vers un maximum local (Baum et Welsh).

L'algorithme EM peut donc s'écrire ainsi :

- Initialisation au hasard de $\theta^{(0)}$ (si l'on ne dispose pas d'un $\theta^{(0)}$ plus satisfaisant à l'initialisation)
- $c = 0$
- Tant que l'algorithme ne converge pas
  - On évalue l'espérance $Q(\theta;\theta^{(c)})$, c'est l'étape E.
  - On maximise $\theta^{(c+1)}$, c'est l'étape M.
  - $c = c + 1$

- Fin.

Nous avons présenté ici la version la plus simple de l'algorithme EM, qui est celle que nous utilisons. Il existe des variantes de cet algorithme (comme GEM de [Dempster *et al.*, 1977], CEM de [Celeux & Govaert, 1991] ou encore SEM de [Celeux, 1985]).

L'avantage du clustering flou est qu'en fournissant une probabilité d'appartenance à chaque cluster pour chaque phrase, il nous fournit un score entre 0 et 1 qu'il est bien plus facile de combiner avec les scores de transition et / ou de rupture présentés dans la

section précédente. Ainsi, une hybridation des deux approches est facilitée [32]Les hybridations testées seront présentées dans le chapitre suivant.).

3.7 Conclusion

Nous concluons donc ce chapitre consacré aux aspects les plus théoriques de notre travail.
Dans la section 3.2, nous avons présenté les différentes possibilités qui se sont offertes à nous en matière de représentation du texte, et notamment en matière de représentation vectorielle de texte. Le choix des vecteurs sémantiques apparaît clairement comme le plus adapté à la fois à notre tâche de segmentation thématique et à notre objectif d'éloignement vis à vis du lexique.
En section 3.3 nous avons précisé la notion de segment thématique telle que nous la percevons. En effet, la littérature ne donne pas de définition claire de ce qu'est un segment thématique. En palliant ce manque nous dessinons clairement les contours de ce que nous souhaitons obtenir comme résultat de la tâche de segmentation thématique.
Dans les sections 3.5 et 3.6, nous avons mis en avant les différentes options que nous avons envisagées pour venir à bout de notre tâche de segmentation thématique en différentiant clairement la détection active des frontières (section 3.5) de la détection passive des frontières (section 3.6). Le choix des différentes approches que nous avons effectués dans ces deux sections découlent directement de ceux effectués en matière de représentation du texte et de définition du segment thématique. Nous conservons donc une cohérence entre la représentation du texte et la tâche de segmentation thématique que nous nous sommes fixée.
Mais l'adéquation entre la théorie et la pratique ne se vérifie pas toujours, et afin de pouvoir vérifier la validité de nos théories il est nécessaire de disposer d'outils nous permettant de les tester. Le chapitre suivant sera donc consacré à la mise en pratique du cadre théorique que nous avons défini ici.

32. (

4

Mise en place d'une application de segmentation thématique de texte : Transeg

Sommaire

4.1 Introduction

Afin de valider les assomptions, très théoriques, faites dans le chapitre précédent, il nous était nécessaire de disposer d'une application automatique les mettant en œuvre.
Ce chapitre s'attache à décrire **Transeg** (pour Transition & Segmentation), l'application que nous avons développée afin de pouvoir vérifier si nos hypothèses sur l'organisation thématique d'un texte étaient valides, et, le cas échéant, s'il est possible d'hybrider notre approche avec d'autres approches.
La section 4.2 de ce chapitre va s'attacher à décrire globalement les différentes étapes qui composent la segmentation thématique par Transeg. Transeg se veut une application la plus modulaire possible. De ce fait chaque étape est aussi indépendante que possible des autres.
La section 4.3 présente la manière dont les vecteurs sémantiques sont obtenus grâce à

l'application SYGFRAN ([Chauché, 2001]). Cette section détaille donc le fonctionnement de SYGMART et son application spécifique d'analyse de la langue française SYGFRAN, en s'intéressant tout particulièrement à la manière dont les vecteurs sémantiques sont générés. On notera que, toujours dans un souci de modularité, Transeg peut utiliser toute autre représentation vectorielle du texte. SYGFRAN étant un outil performant et développé au sein de notre équipe, c'est sur ses vecteurs sémantiques que notre choix s'est porté.
La section 4.4 décrit en détail les outils que nous avons utilisés pour comparer les vecteurs sémantiques. Nous y étudions les différentes possibilités qui se sont présentées à nous en matière de distance et de similarité, pour y présenter finalement notre distance de concordance.
La section 4.5 est une description plus générale de l'implémentation des différentes phases de Transeg, ainsi que des algorithmes testés lors du développement de Transeg.
Nous conclurons ce chapitre dans la section 4.6.

4.2 Architecture

Cette section décrit brièvement le fonctionnement global de Transeg en présentant son architecture. Ainsi, la segmentation d'un texte par Transeg s'articule en trois grandes phases que nous décrivons ici et qui sont résumées dans la figure 4.1. La première phase est correspond à la mise en forme du texte et à sa transformation en une succession de vecteurs sémantiques. La phase 2 utilise les vecteurs sémantiques pour retrouver les zones de transitions et les frontières thématiques. La dernière phase quant à elle met en forme le résultat.
Nous présentons dans la suite chacune de ces phases de manière générale pour donner une vision d'ensemble de l'application. Les étapes du développement de Transeg présentant un intérêt particulier seront détaillées dans les sections suivantes.

4.2.1 Première phase : Génération des vecteurs sémantiques

Le texte à segmenter présenté en entrée de l'application Transeg est tout d'abord segmenté en phrases et balisé afin que chaque phrase ressemble à la phrase donnée en exemple dans la figure 4.2.
Chaque phrase est donc identifiée par une balise qui marque son début. Cette balise contient deux informations essentielles pour la suite de l'analyse du texte :

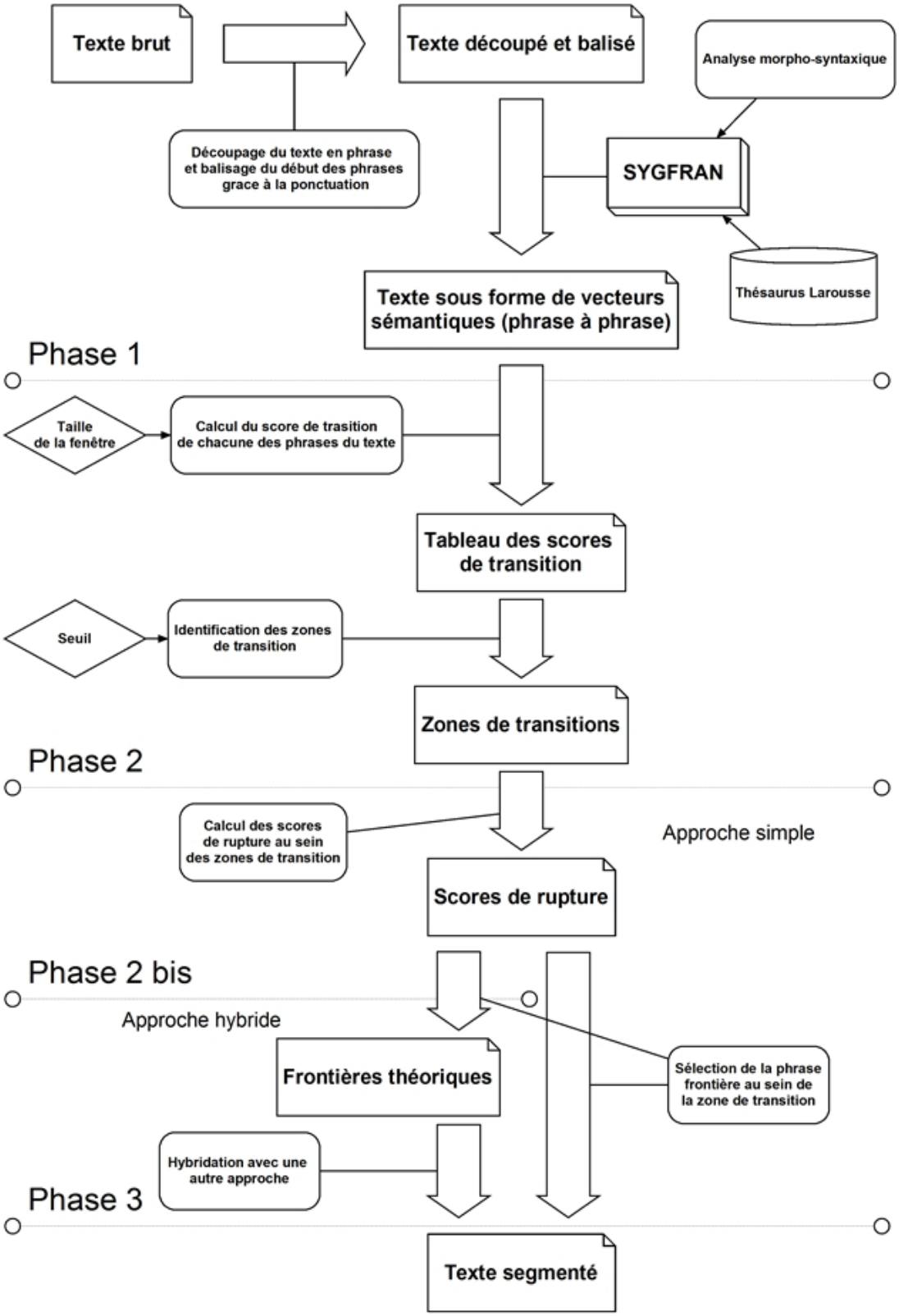

FIGURE 4.1 – Architecture globale de Transeg

– Le numéro de la phrase qui nous permettra d'identifier plus facilement les phrases par la suite.

<1x> Avec l'acquisition de la branche basse tension de la filiale du groupe Daimler, General Electric renforce ses positions en Europe où il réalise déjà 15 % de son chiffre d'affaires.

FIGURE 4.2 – Exemple de phrase après segmentation en phrase et balisage

– La phrase est-elle ou non le début d'un segment thématique. Cette information est matérialisée par le x au sein de la balise. On notera que cette information n'est jamais disponible pour les différents algorithmes de segmentation thématique utilisés par la suite, elle n'est utile que pour évaluer les résultats à la fin.

Le balisage des phrases se fait automatiquement, mais la segmentation en phrases est le résultat d'un processus semi-automatique. Le texte est donc d'abord segmenté en phrases automatiquement en utilisant des heuristiques simples, traduites en expressions régulières, puis corrigé si nécessaire par un opérateur humain. Ce choix a été fait pour trouver un juste équilibre entre complexité d'implémentation et vitesse. Notons que la taille raisonnable du corpus (cinquante textes) est également à l'origine de ce choix.
Une fois le texte correctement préparé et mis en forme, il est envoyé phrase à phrase à l'analyseur SYGFRAN. SYGFRAN retourne le vecteur sémantique de chaque phrase sous la forme d'un tableau de 873 flottants. Les tableaux de chacune des phrases seront stockés, les uns à la suite des autres, dans un fichier dit de vecteurs sémantiques. Ce fichier sera par la suite utilisé par Transeg pour effectuer les différents calculs nécessaires à la segmentation.

4.2.2 Deuxième phase : Identification des zones de transition

L'identification des zones de transition se fait ensuite en travaillant sur le fichier contenant les vecteurs sémantiques du texte, ainsi que le fichier source (le fichier texte une fois préparé par la phase 1). Le véritable travail se fait sur le fichier des vecteurs sémantiques, le fichier texte correspondant n'étant là que pour la génération du résultat, une fois la segmentation achevée.
L'utilisateur choisit la taille de la fenêtre ainsi que le seuil de transition et l'algorithme présenté en 3.5 est appliqué au texte. Les phrases successives ayant un score de transition supérieur au seuil entré par l'utilisateur forment les zones de transitions. Les phrases isolées sont elles ignorées.

4.2.3 Troisième phase : Sélection des phrases frontières

Toujours en suivant l'algorithme présenté en 3.5, les scores de transition des phrases identifiées comme faisant partie d'une zone de transition sont modifiés par les scores de rupture. Les phrases frontières sont alors identifiées dans chaque zone de transition.
On notera toutefois que l'éventuelle hybridation avec une méthode de clustering se fait à cette étape. Nous comparons les frontières obtenues par Transeg et celles données par l'algorithme avec lequel nous souhaitons hybrider Transeg et une hybridation s'effectue en fonction de l'algorithme utilisé. Il s'agit donc là, plus d'un système de vote que d'une véritable hybridation.
Une fois les frontières déterminées, les résultats sont présentés de trois manières différentes :

- Un fichier ne contenant que les en-tête des phrases. Ce fichier sert pour l'évaluation des résultats principalement.
- Un fichier texte du type de celui que nous obtenons après la phase 1, mais avec des sauts de ligne pour séparer les segments thématiques. Ce fichier s'adresse surtout à un éventuel utilisateur souhaitant voir le résultat.
- Un fichier HTML qui servira pour l'évaluation humaine sur Internet (voir chapitre 5).

4.3 Les vecteurs sémantiques par SYGFRAN

Dans cette section, nous présentons l'analyseur de la langue française SYGFRAN qui nous fournit les vecteurs sémantiques que nous utilisons dans nos travaux.
SYGFRAN est en réalité une application développée sur la base d'un outil de manipulation d'éléments structuré SYGMART présenté par [Chauché, 1984]. Nous présenterons brièvement SYGMART afin d'aider à la compréhension de SYGFRAN, puis nous nous attacherons à étudier comment sont générés les vecteurs sémantiques par SYGFRAN. Des informations plus détaillées sur le fonctionnement de SYGMART et SYGFRAN sont disponibles dans le manuel de référence de l'application [Chauché, 2001].

4.3.1 SYGMART

SYGMART (pour Système Grammatical de Manipulation Algorithmique et Récursive de Texte) est un environnement opérationnel permettant la décomposition et la transformation en arbre d'une chaîne textuelle, puis la manipulation de cette structure arborescente pour enfin re-linéariser la chaîne textuelle.
SYGMART se compose de trois modules distincts : OPALE, TELESI [33] et AGATE. Chacun de ces modules effectue un traitement en se basant sur un ensemble de règles et un dictionnaire.

4.3.1.1 OPALE : le module de décomposition morphologique

L'objectif du module OPALE est de convertir une chaîne passée en entrée (la chaîne à analyser) en une décomposition morphologique. OPALE définit une transition entre une chaîne textuelle et une structure de type arborescente correspondante. En s'appuyant sur un dictionnaire et la segmentation de la chaîne à analyser [34], le module construit un transducteur d'états finis qui est utilisé pour effectuer cette transition.
La figure 4.3 est un très court exemple de la décomposition de la chaîne *« au marché »* par le module OPALE pour l'application SYGFRAN. Les nœuds sont étiquetés par une numérotation effectuée sur un parcours en profondeur. La racine ancre le texte passé en entrée (une phrase, un paragraphe ou, dans cet exemple, un morceau de phrase). Les nœuds fils de la racine ancrent chacun des mots du texte. Les nœuds descendants des

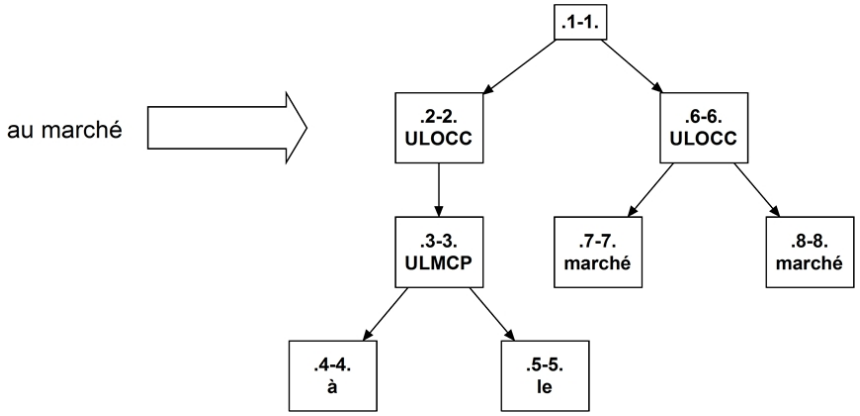

FIGURE 4.3 – Exemple de décomposition par OPALE pour SYGFRAN

33. TELESI (Transduction d'ÉLÉments Structurés Indexés) en référence au terme *télésie* du grec « parfait » qui fut utilisé par Haüy pour désigner les trois gemmes les plus précieuses d'Orient, à savoir le rubis,le saphir et la topaze d'Orient. Les noms des deux autres modules font références eux aussi au pierres précieuses.
34. Une chaîne textuelle est segmentée en unités syntaxiques de taille minimum (mots, symboles de ponctuation, etc.)

nœuds associés à un mot ancrent une solution de segmentation possible pour ce mot. Pour les mots complexes, ces nœuds peuvent avoir plusieurs fils et donc plusieurs solutions possibles.
Dans notre exemple, le mot *au* a été segmenté en une forme non contractée *à le*, mais reste toutefois un mot simple. Dans le cas de *marché*, par contre, OPALE nous propose deux solutions : une pour le nom commun *marché* (le lieu d'échange et de commerce) et une pour la forme participe passé *marché* du verbe marcher.
Les problèmes d'ambiguïté tels que celui qui apparaît sur le mot *marché* ne pouvant pas être réglés au niveau morphologique, ils seront traités lors de l'analyse syntaxique par le module TELESI.

4.3.1.2 TELESI : le module de transformation d'éléments structurés

TELESI est un module définissant des transitions entre différents éléments structurés. Ainsi si (E, S) est un élément structuré, avec :

- E un ensemble fini d'étiquettes.
- S un ensemble d'arborescences étiquetées (A_i, E_i, f_i) tel que $E_i \in E$ et avec A_i une arborescence et f_i une fonction qui associe à chaque nœud de A_i un et un seul élément de E_i.

TELESI prend en entrée soit la sortie d'OPALE, soit la sortie d'un autre traitement TELESI et transforme la structure arborescente passée en entrée en une autre structure arborescente. Cette transformation est faite à l'aide d'un transducteur à pile composé, simulant une grammaire transformationnelle.
TELESI procède par transformations successives pour arriver au résultat voulu. Ainsi, si nous prenons la phrase *« Le chat mange la souris blanche. »*.
La sortie sortie OPALE donnera un résultat proche de celui présenté par la figure 4.4

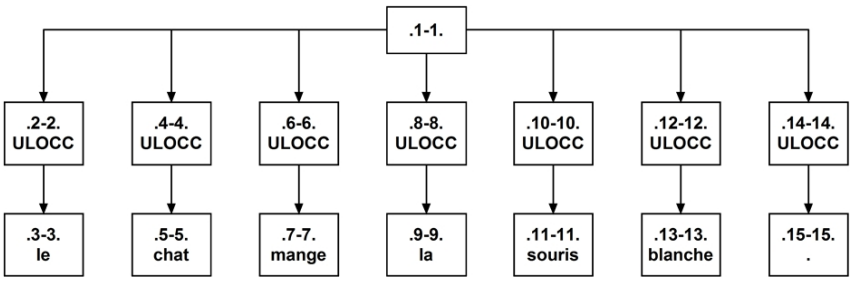

FIGURE 4.4 – Sortie OPALE pour la phrase : *« Le chat mange la souris blanche. »*

(nous avons simplifié la sortie pour des raisons de lisibilité). Or ce résultat est pour le

moment « plat », c'est à dire que l'arbre n'incorpore pas encore les relations hiérarchiques entre les différents constituants de la phrase.
La sortie finale de TELESI (celle qui sera fournie à l'utilisateur, ou à OPALE), ressemblera elle à celle qui est présentée dans la figure 4.5.

Nous voyons clairement dans la figure 4.5 que *la souris blanche*, complément d'objet

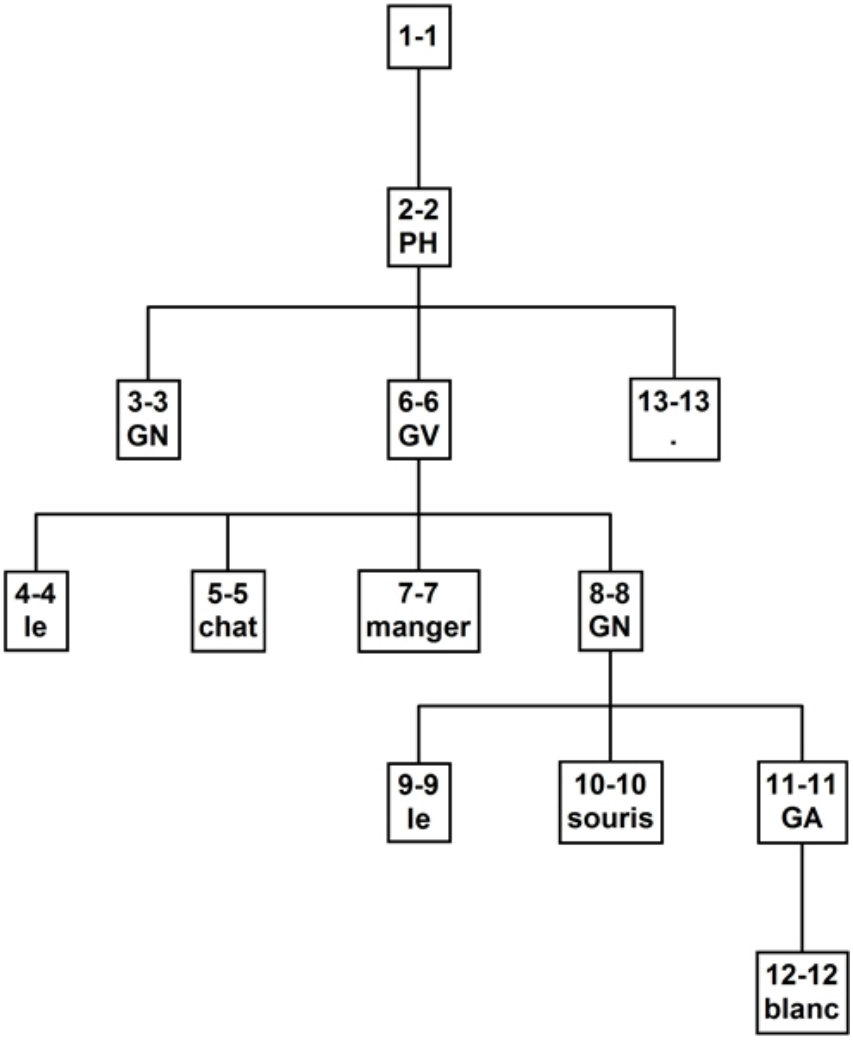

FIGURE 4.5 – Exemple de sortie du module TELESI pour SYGFRAN : *Le chat mange la souris blanche*

direct du verbe manger, a été inclus dans le groupe verbal et donc, en quelque sorte « inféodé » au verbe.
Nous notons également qu'à la numérotation attribuée aux nœuds par OPALE, TELESI a ajouté :

– La catégorie lexicale du constituant pour les nœuds internes. Ainsi, la racine de l'arbre se voit attribuer l'étiquette PH pour indiquer que l'ensemble de la chaîne est une phrase. On voit également apparaître les étiquettes GN, GV et GA pour

respectivement groupe nominal, groupe verbal et groupe adjectival. Les catégories absentes de l'exemple ayant, bien entendu, aussi des étiquettes pour les identifier.
– Le lemme pour les feuilles. Ainsi *mange* devient *manger*, *la* devient *le* et *blanche* devient *blanc*.

Le système TELESI peut être vu comme un réseau conditionnel de grammaires élémentaires. TELESI ne gère donc pas une grammaire unique, mais un ensemble de petites grammaires inter-connectées. Dés lors, le traitement d'un élément structuré spécifique devient un cheminement dans ce réseau avec, en chaque point, l'application d'une grammaire élémentaire. Le réseau étant conditionnel, le chemin l'est également et dépend donc de l'élément structuré traité. Parmi les grammaires du réseau, certaines sont définies comme initiales et peuvent donc servir de point de départ au cheminement (toujours en fonction de l'élément structuré traité). Il en va de même pour les sorties du réseau qui sont identifiées par des marqueurs. TELESI ne considérera une transition comme terminée que lorsque l'un de ces marqueurs est atteint.
Nous ne rentrerons pas plus dans le détail pour ce qui est des grammaires TELESI, leur fonctionnement est décrit plus en détail dans le manuel d'utilisateur [Chauché, 2001] ou encore dans la thèse de Mehdi Yousfi-Monod [YM2007].

4.3.1.3 AGATE : le module de linéarisation d'éléments structurés

AGATE définit une transition entre un élément structuré (un arbre) et une chaîne de caractères. Tout comme OPALE, AGATE est un transducteur d'états finis non déterministe qui s'appuie sur les étiquettes de chacun des nœuds de la structure arborescente pour fournir une solution. On notera que contrairement à OPALE, AGATE ne fournit que la première solution possible (OPALE prend en compte toute les solutions possibles).

4.3.2 SYGFRAN

Le système opérationnel SYGMART que nous venons de présenter (brièvement) est l'environnement au sein duquel le programme SYGFRAN est exécuté. SYGFRAN est donc un ensemble de règles (grammaires TELESI) et de dictionnaires (pour OPALE et AGATE) visant à produire une analyse morpho-syntaxique du français. Mais l'intérêt principal de SYGFRAN est de fournir des vecteurs sémantiques. Ces vecteurs sémantiques sont produits en se basant sur les concepts du thésaurus Larousse ([Larousse, 1992b]) et en prenant en compte les résultats de l'analyse morpho-syntaxique.
Nous allons décrire ici le fonctionnement de SYGFRAN, en nous attachant surtout à la

manière dont l'application génère les vecteurs sémantiques.

4.3.2.1 L'analyse syntaxique dans SYGFRAN

L'analyse syntaxique de SYGFRAN se base sur un ensemble de grammaires TELESI visant à reproduire le plus fidèlement possible la grammaire française.

4.3.2.2 La génération des vecteurs sémantiques

La méthode vectorielle que nous présentons ci-dessous peut être utilisée pour représenter aussi bien un mot qu'un ensemble ordonné de mots[35]. Transeg étant une application de segmentation thématique considérant la phrase comme élément atomique, nous nous attacherons plus particulièrement à l'obtention du vecteur sémantique de phrase.

Vecteur de terme
Définition : Un vecteur sémantique projette un terme donné dans un espace sémantique dont une famille génératrice correspond à un ensemble d'idées. L'ensemble des idées nécessaires pour former une famille génératrice peut être définie par un thésaurus.

La procédure est la suivante : on projette la totalité des lexies du dictionnaire sur un espace défini à partir d'une famille de concepts « à la Roget » ([Roget, 1852]). Le thésaurus proposé par Roget en 1852 définit une famille de 1000 concepts hiérarchisés en 4 niveaux, très centrée autour des notions de métaphysique et de religion. Cette orientation, sans doute liée aux préoccupations et sujets importants de l'époque, est aujourd'hui critiquée car elle fausserait les analyses. Malgré cela, le Roget reste encore très largement utilisé dans la communauté du TALN s'intéressant à la langue anglaise.
Pour le Français, les lexicologues du Larousse ([Larousse, 1992b]) ont défini une famille de 873 concepts hiérarchisés en 4 niveaux sur le modèle de Roget. Sur un plan vectoriel, cela produit un espace à 873 dimensions que l'on admet comme étant de dimension donnée. Les approches « à la Roget » sont relativement nombreuses depuis quelques années, dans la littérature anglo-saxonne. En Français, l'indexation automatique à partir du thésaurus a été proposée à l'origine par Jacques Chauché, mais on la retrouve aujourd'hui utilisée dans d'autres travaux.
Formellement, on considère que tout terme t du dictionnaire est représenté par un vecteur $\vec{t}$ dans l'espace vectoriel considéré, que l'on nommera $\vec{\mathcal{V}}$. On suppose qu'il existe une

35. Que ce soit une proposition, une phrase ou même un texte.

application qui plonge l'espace lexical linguistique dans l'espace vectoriel engendré par la famille de concepts du thésaurus. Pour des besoins de calcul, seule une version normée $\vec{t}_{nor}$ de ce vecteur est conservée dans l'espace. Comme on ne traite que de vecteurs normés, par convention, on écrira $\vec{t}$ pour désigner le vecteur normé du terme t.
Pour cela, on introduit une norme euclidienne sur l'espace vectoriel sémantique.
La majorité des mots, étant polysémique, renvoie à une multiplicité d'idées, ou concepts du thésaurus.

Exemple

Les concepts associées au mot *<calcul>* sont par exemple : CALCUL, OPERATION ARITHMETIQUE, MALADIE et INTENTION. L'emploi de ce mot simplement ne permet donc pas de définir sa signification : par exemple, *calcul arithmétique* ou *calcul biliaire*, ou *Il m'a aidé par calcul*.
Cela signifie que le terme doit être représenté, non seulement par la manière dont il est indexé dans le thésaurus, mais aussi par ses différentes significations, qui elles, n'ont de sens que lorsque le mot est utilisé dans une construction (groupe ou phrase). Le calcul sémantique sur une phrase doit donc incliner le sens du mot *<calcul>* vers une des significations possibles.

Vecteur sémantique d'une phrase

Définition : On dira que l'on représente toute *phrase* construite, par un vecteur obtenu à partie d'une combinaison linéaire de vecteurs sémantiques des *groupes* qui la composent. On dira que l'on représente tout *groupe* construit, par un vecteur obtenu à partie d'une combinaison linéaire de vecteurs sémantiques des *termes* qui le composent.
Pour cela on introduit les opérations suivantes :
Somme normée : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ représentant les vecteurs (normés) de deux termes t_1 et t_2.

$$\overrightarrow{(t_1 + t_2)_{nor}} = \frac{\vec{t_1}+\vec{t_2}}{\|\vec{t_1}+\vec{t_2}\|} \tag{4.1}$$

Remarque : La somme normée n'est pas associative :
$\overrightarrow{(t_1 + t_2 + t_3)_{nor}}$ n'est pas égal à $(\overrightarrow{(t_1 + t_2)_{nor}} + \vec{t_3}))_{nor}$. Par convention, on ne retiendra comme opération de somme **que** la somme normée, et on omettra dorénavant l'indice *nor*.

Multiplication par un scalaire :
Soit un vecteur $\vec{t}$ normé. Soit λ un scalaire. Le vecteur $\vec{\lambda t}$ est égal à $\lambda * \vec{t}$. Cela signifie que toutes les composantes du vecteur sont multipliées par le scalaire.
Remarque : Cette multiplication a pour objectif de renforcer la « présence » du vecteur dans une combinaison linéaire, et ne s'utilise en principe jamais isolément.

Produit terme à terme : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ normés. Le produit terme à terme des deux vecteurs se définit comme :

$$\overrightarrow{(t_1 * t_2)_{nor}} = \frac{\vec{t_1} * \vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{4.2}$$

où si $a_{p,i}$ est la $i^{ème}$ composante de $\vec{t_1} * \vec{t_2}$, et $a_{1,i}$ et $a_{2,i}$ respectivement celles de $\vec{t_1}$, et $\vec{t_2}$, on a :

$$\forall i \in [1, 873], a_{p,i} = a_{1,i} * a_{2,i} \tag{4.3}$$

Par convention, on omettra l'indice *nor* et on appellera par défaut $\overrightarrow{(t_1 * t_2)}$ le produit terme à terme normé.

Distance « angulaire » : La distance selon Salton ([Salton *et al.*, 1975]), servant de mesure de similarité est calculée comme le *cosinus* de l'angle de deux vecteurs.

$$sim(\vec{t_1}, \vec{t_2}) = \cos \widehat{\vec{t_1}, \vec{t_2}} = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{4.4}$$

où « . » est le produit vectoriel classiquement défini.

La distance que nous utilisons correspond à une mesure relative à l'angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l'espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

$$\delta(\vec{t_1}, \vec{t_2}) = 1 - \cos \widehat{\vec{t_1}, \vec{t_2}} \tag{4.5}$$

Remarques : Ramener les valeurs de δ à $[0, 1]$ est plus pratique que de mesurer des valeurs entre 0 et $1,67$ radians. Lorsque deux vecteurs sont totalement divergents (intersection vide), leur angle est de $\frac{\pi}{2}$, et le cosinus vaut 0 : leur distance est maximale et vaut 1. Lorsque ces vecteurs sont très proches, leur angle tend vers 0 , le cosinus tend vers 1 et la distance, vers 0. On notera que la distance angulaire ne vérifie les trois propriété d'une distance (et donc est réellement une distance) que si nous ne travaillons qu'avec des valeurs positive ou nulle dans nos vecteurs. Ce qui est le cas, aucun concept ne pouvant être activé négativement dans le thésaurus Larousse.

Tous les vecteurs ont un angle forcément compris entre 0 et $\frac{\pi}{2}$, par construction, et appartiennent au même espace vectoriel.

Vecteur de groupe La deuxième propriété du calcul sémantique correspond à une définition différentiée d'un groupe suivant sa structure. Ainsi, le sens du groupe « le calcul du sens » est distinct du sens du groupe « le sens du calcul »,ces deux groupes ayant

rigoureusement les mêmes éléments (le langage naturel n'étant pas commutatif). Comme le mot « sens » est très riche sémantiquement (une vingtaine de sens justement) nous prendrons pour l'exemple de la représentation l'idée associée : Sens. L'idée est différente du terme, selon les lexicologues, en ce qu'elle étiquette un champ sémantique. Le terme peut appartenir ou relever de plusieurs champs, en raison de sa polysémie.
Dans le sous-espace ayant comme axe *Calcul*, *Intention* et *Sens* les vecteurs associés aux deux groupes précédents pourront être visualisés comme présenté dans la figure 4.6.

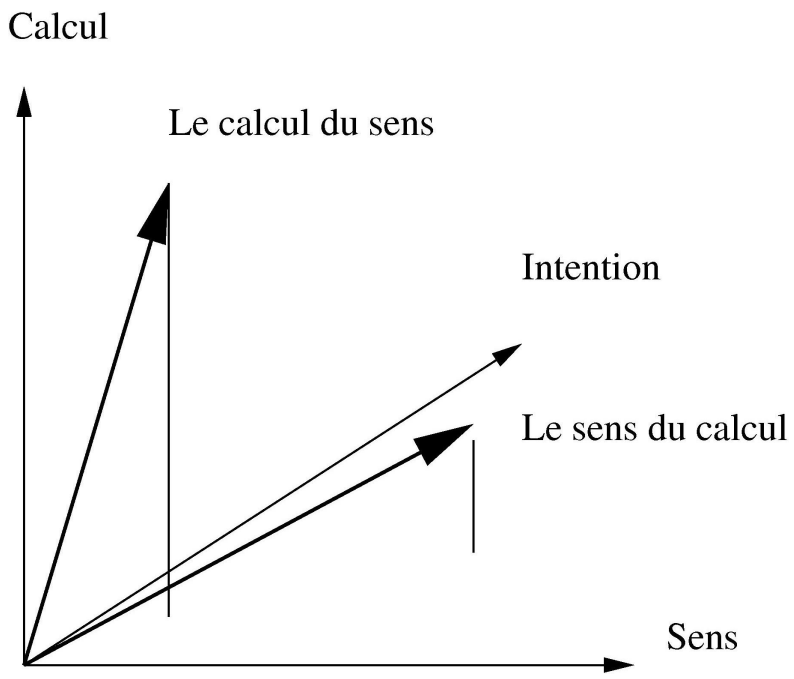

FIGURE 4.6 – Vecteur de groupe

Calcul du vecteur de phrase Le calcul d'un vecteur de phrase s'effectue (sur une phrase) en plusieurs étapes à partir de la structure syntaxique :

- La première étape consiste à associer à chaque feuille un vecteur sémantique issu de la lecture d'un dictionnaire (vecteur génératif).

Si un élément a plusieurs sens ou interprétations possibles, le vecteur associé correspond au *centroïde* de l'ensemble des vecteurs associés à chaque interprétation (somme normée de tous les vecteurs indexant ce terme).

- La deuxième étape consiste à calculer récursivement le vecteur associé à chaque groupe.

Le vecteur associé à un groupe est obtenu par une combinaison linéaire des vecteurs associés aux éléments de ce groupe. Les cœfficients de cette combinaison linéaire dépendent de la fonction syntaxique de l'élément : gouverneur du groupe, sujet, objet, etc.

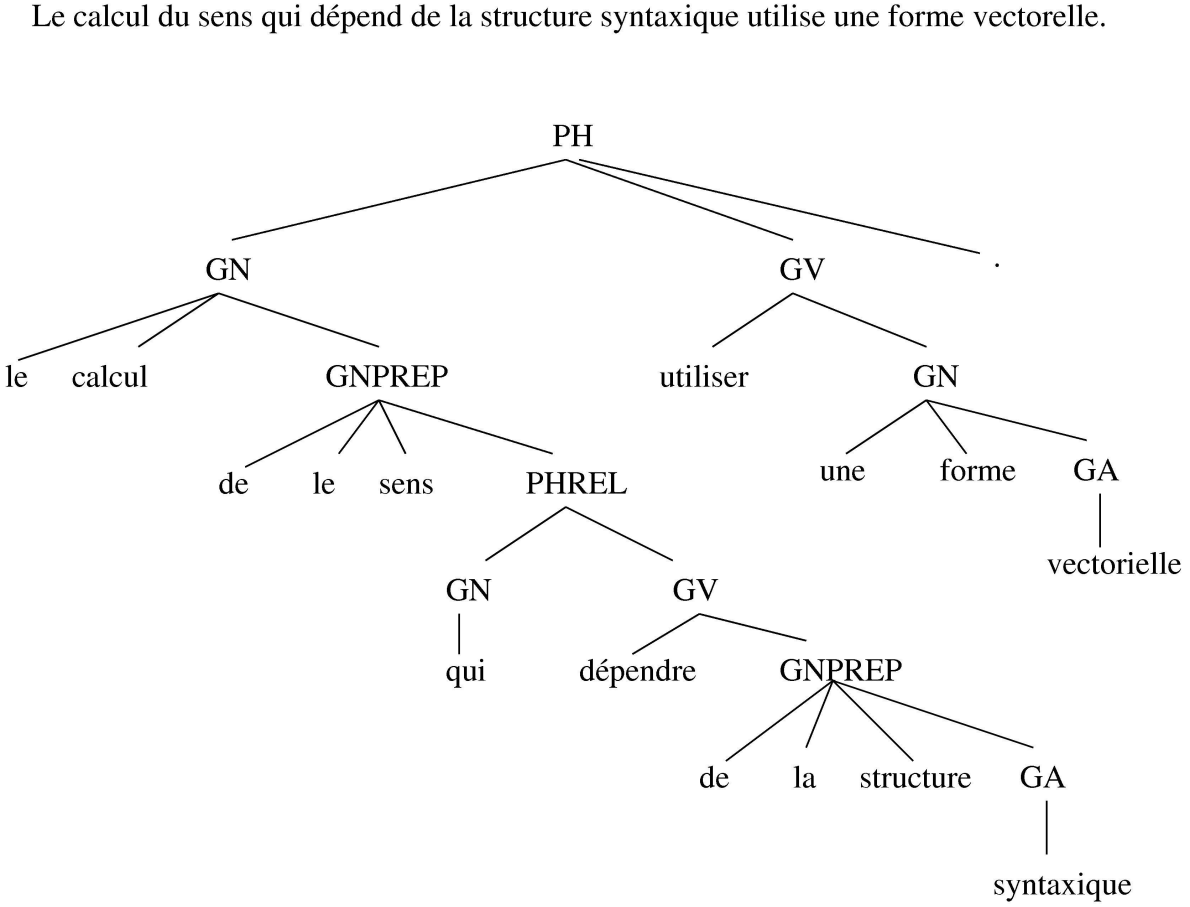

FIGURE 4.7 – Structure syntaxique

– La troisième étape actualise les vecteurs associés aux feuilles. Cette actualisation consiste à effectuer un produit terme à terme entre le vecteur à actualiser et le vecteur obtenu du texte.

Cette actualisation terminée un nouveau calcul est effectué. La convergence est très rapide et deux itérations suffisent pour obtenir un vecteur significatif.

4.4 Le choix de l'outil de comparaison des vecteurs sémantiques

Dans la mesure où nous travaillons avec des vecteurs sémantiques, dans l'espace vectoriel défini des 873 concepts du thésaurus Larousse ([Larousse, 1992b]), il fallait nous doter d'outils permettant de comparer efficacement ces vecteurs. Nous explorons ici les différentes possibilités qui se sont offertes à nous, et présentons celles que nous avons choisi d'exploiter.
Comme nous sommes dans une tâche de segmentation thématique, notre outil de comparaison doit se focaliser sur ce qui différentie ou rapproche thématiquement deux phrases. Ce qui nous amène à nous intéresser aux notions de distance et de similarité.

4.4.1 Distance et similarité en segmentation thématique

4.4.1.1 De la distance mathématique à la distance thématique

« En mathématiques, une distance est une application qui formalise l'idée intuitive de distance, c'est à dire la longueur qui sépare deux points. » [Larousse, 1992a].
Plus précisément, c'est une application $d : E \times E \rightarrow \Re+$ qui vérifie les trois propriétés suivantes :

- La symétrie $\forall x, y \in E, d(x, y) = d(y, x)$.
- La séparation $\forall x, y \in E, d(x, y) = 0 \Leftrightarrow x = y$.
- L'inégalité triangulaire $\forall x, y, z \in E, d(x, z) \leq d(x, y) + d(y, z)$.

La distance mathématique est particulièrement bien adaptée à notre représentation du texte. En effet, nous nous appuyons sur une représentation vectorielle en vecteur sémantique. Ils nous est donc aisé de considérer ces vecteurs comme les coordonnées des points issues de la translation du point d'origine de notre espace vectoriel par les vecteurs sémantiques.
Comme nous nous situons dans une tâche de segmentation thématique, nous utiliserons par la suite le raccourci de langage « distance thématique » pour désigner la mesure de l'éloignement thématique d'une phrase par rapport à une autre, sur la base de notre représentation vectorielle. En vérité, la notion de distance thématique s'écartera un peu du cadre rigide de la définition mathématique de la notion de distance. Notamment, l'inégalité triangulaire ne sera pas toujours vérifiée sur l'ensemble de départ de la fonction de distance, mais sera toujours vérifié sur notre ensemble de travail.

4.4.1.2 La similarité

Contrairement à la distance, la similarité n'est pas un concept clairement défini du point de vue mathématique. La similarité est un score entre deux phrases qui, au contraire de la distance, sera plus élevé si les deux phrases sont proches.

4.4.2 La distance euclidienne

La distance euclidienne est une généralisation du théorème de Pythagore à des espaces vectoriels à n dimensions (des espaces euclidiens). C'est une des plus anciennes et des plus simples distances entre points, et elle est encore très utilisée, sous sa forme originale, ou sous sa forme généralisée la p-distance.
Soit deux points x_i et y_i de respectivement $E, (x_1, x_2, ..., x_i, ..., x_n)$ et $E, (y_1, y_2, ..., y_i, ..., y_n)$, on a :

$$p - distance = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{4.6}$$

La distance euclidienne étant la 2-distance, on a la distance euclidienne :

$$D_{euc} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.7}$$

Le problème est que la distance euclidienne est une distance « ouverte ». En effet, c'est une application $d : E \times E \rightarrow \Re+$, $\Re+$ ayant une borne inférieure (0), mais pas de borne supérieure (par définition $\Re+ = [0, +\infty[$). Ainsi la distance euclidienne peut nous permettre de comparer les vecteurs sémantiques relativement les uns aux autres. En effet, on peut aisément dire si un vecteur sémantique est « plus loin » ou « plus près » d'un autre qu'un troisième. Mais la distance euclidienne nous handicape sérieusement lorsqu'il s'agit de comparaison absolue. Sans borne maximale, il nous est impossible de déterminer un seuil à partir duquel nous considérons qu'un vecteur sémantique est « thématiquement éloigné ».
Cette absence de borne supérieure nous a fait écarter la distance euclidienne comme distance thématique.

4.4.3 Le cosinus

Le cosinus est sans conteste la mesure de similarité la plus utilisée en TALN, elle fut notamment choisie comme mesure de similarité de référence par Salton ([Salton *et al.*, 1975]) dans son modèle vectoriel. Ainsi, si l'on considère deux phrases p_1 et p_2 représentées par leur vecteurs sémantiques respectifs $V(p_1)$ et $V(p_2)$, alors la « *similarité cosinus* » $sim(p_2, p_2)$ des deux phrases sera :

$$sim(p_2, p_2) = \frac{V(p_1).V(p_2)}{\|V(p_1)\| \, . \, \|V(p_2)\|} \tag{4.8}$$

4.4.4 La distance angulaire

Afin de pouvoir mesurer la différence thématique entre deux phrases, deux centroïdes ou encore entre une phrase et un centroïde il nous faut disposer d'une fonction similarité ou d'une distance. Une des distances que nous avons choisi de tester est la distance thématique présentée par [Lafourcade & Prince, 2001]. Ainsi, si X et Y sont deux vecteurs, D_A étant la distance thématique recherchée, on a :

$$D_{ang} = \arccos(\cos(\widehat{X, Y})) \tag{4.9}$$

La distance D_A étant exprimée ici en radians.
Le principal avantage de la distance angulaire sur le cosinus (dont elle est très proche), en dehors d'être une distance mathématique sur notre ensemble de travail, est son comportement entre 0 et $\frac{\pi}{4}$. LŠ*arccosinus* est une fonction fortement non linéaire pour des valeurs dŠangles faibles (inférieur à $\frac{\pi}{4}$, soit inférieur à 45ř), alors quŠelle se comporte de manière quasi linéaire pour les valeurs dŠangles élevées. Ainsi on obtient une plus grande finesse dŠanalyse lorsque deux phrases sont sémantiquement proches. On notera que le calcul de la distance angulaire que nous proposons fait appel à une « astuce » permettant de la retrouver à partir du calcul plutôt simple du cosinus. Ainsi, la formule développée de la distance angulaire devrait être :

$$D_{ang} = \arccos(\frac{X.Y}{\|X\| \, . \, \|Y\|}) \tag{4.10}$$

4.4.5 La distance de concordance

Les vecteurs sémantiques issus de l'analyse par SYGFRAN ont 873 composantes, et la grande majorité de ces dernières ne sont pas activées pour une phrase donnée. Avec autant de valeurs nulles, les méthodes classiques de comparaison entre deux vecteurs, telles que celles présentées plus haut, sont trop peu discriminantes pour permettre une analyse fine

du texte.

4.4.5.1 Environnement et automobile : un exemple des limites des mesures « classiques »

Traitons un exemple simplifié pour voir plus en détail les faiblesses des mesures de similarité ou de distance présentée plus haut.
Prenons par exemple ces deux phrases :

- «*L'impact désastreux de l'automobile sur l'environnement, notamment sur le réchauffement climatique, n'est plus à démontrer.*»
- «*Notre nouveau modèle d'automobile ne rejette que* $127g$ *de* CO_2 *au kilomètre, faites un geste pour l'environnement, achetez une automobile X.*»

Ces deux phrases traitent clairement d'automobile et d'environnement. Mais si la première fait visiblement partie d'un texte généraliste sur les différents facteurs du réchauffement climatique, la seconde est à l'évidence extraite d'une publicité pour un modèle d'automobile.
Peut on affirmer que ces deux phrases appartiennent au même thème ? Le choix est difficile. Encore une fois, nous nous retrouvons face au problème de la subjectivité de la notion de thème. Toutefois, si ces deux phrases abordent des sujets très proches, il existe une différence fondamentale entre les deux. La première a pour thème central l'environnement, l'automobile étant là pour illustrer l'impact de l'activité humaine sur l'environnement (du moins peut on le supposer, puisque nous n'avons pas le contexte). La seconde phrase tourne autour de l'automobile, et même d'un modèle précis d'automobile, l'environnement n'étant alors qu'un argument de vente. Elles abordent donc des concepts similaires, mais des thèmes différents.
Si l'on devait représenter ces deux phrases par deux vecteurs sémantiques on obtiendrait un résultat ressemblant à cela :

- $\vec{Ph}_1 = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0]$ pour la première.
- $\vec{Ph}_2 = [0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0]$ pour la deuxième.

En supposant, bien entendu, que la sixième composante des vecteurs corresponde au concept d'automobile et la onzième à celui d'environnement [36]. Si nous calculons le cosinus de ces deux vecteurs nous obtenons :

36. Cette représentation a été fortement simplifiée pour des questions de lisibilité.

$$cos(V_1, V_2) = \frac{P\vec{h}_1 \times P\vec{h}_2}{\left\|P\vec{h}_2\right\| \times \left\|P\vec{h}_2\right\|} = \frac{4}{5} \tag{4.11}$$

En considérant, comme c'est souvent le cas dans la littérature, le cosinus comme une similarité, nous obtenons une similarité de $\frac{4}{5}$. Cela implique que si l'on considère un cosinus de 1 comme une similarité de 100%, c'est deux phrases sont similaire à 80% ($\frac{4}{5} = 0,8$).
De la même manière, nous obtenons une distance angulaire de :

$$D_{ang}(P\vec{h}_1, P\vec{h}_2) = arccos(cos(P\vec{h}_1, P\vec{h}_2)) = arccos(\frac{4}{5}) \approx 0,64 \tag{4.12}$$

$0,64$ radians, ce qui implique une échelle de mesure allant de 0 à $\frac{\pi}{2}$. Si nous normalisons par $\frac{\pi}{2}$ pour que notre ensemble d'arrivée soit $[0, 1]$, cela ramène notre distance à environ $0,41$. Ce qui est une distance pouvant prêter à confusion. En effet, c'est assez « éloigné » pour dire que les phrases sont différente, mais pas suffisamment pour trancher définitivement la question de savoir si elles appartiennent au même thème ou pas. Ce qui corrobore notre précédente remarque sur la distance angulaire, comme quoi cette dernière est plus discriminante que le cosinus.
Enfin, la distance euclidienne appliquée sur le même exemple donne :

$$D_{euc}(P\vec{h}_1, P\vec{h}_2) = \sqrt{2} \tag{4.13}$$

Ce qui sur une distance ouverte comme la distance euclidienne est très faible et donc dénote une grande proximité entre les deux phrases.

4.4.5.2 Principes de la distance de concordance

L'idée de départ de la mesure de concordance, puis de son évolution, la distance de concordance, est que les valeurs respectives des composantes du vecteur représentant la phrase ne sont pas l'unique source d'information dont nous disposons.
On notera que la mesure de concordance présentée ici n'a rien à voir avec la mesure de concordance proposée par le mathématicien Kendall à la fin des années 40. La similarité entre les noms est involontaire, le terme de concordance ayant été choisi uniquement parce qu'il paraissait être le plus adapté pour décrire cette distance.
Ainsi, si dans l'exemple précédent nous prenons en compte le classement de ces composantes en plus de leur valeur, nous obtenons une nouvelle information. En effet, si nous classons les valeurs nous voyons que le concept d'environnement vient avant celui d'automobile dans la première phrase, et que c'est tout le contraire dans la deuxième phrase.
La mesure, puis la distance, de concordance visent donc à exploiter cette information en

incorporant le rang respectif des différentes composantes du vecteur dans leur calcul.

4.4.5.3 Description de la distance de concordance

La distance de concordance s'appuie sur la mesure de concordance présentée par [Chauché *et al.*, 2003]. Nous l'avons modifiée afin de l'adapter aux besoins de nos travaux.
Considérons deux vecteurs issus du même espace vectoriel $\vec{A}$ et $\vec{B}$. Nous classons leurs composantes de la plus activée à la moins activée et ne conservons que les premières composantes du classement, la quantité conservée est un paramètre Nb à choisir.
$\vec{A_{tr}}$ et $\vec{B_{tr}}$ sont les versions triées et réduites de respectivement $\vec{A}$ et $\vec{B}$. Comme nous ne conservons que les composantes les plus fortes de chaque vecteur, $\vec{A_{tr}}$ et $\vec{B_{tr}}$ peuvent très bien ne pas avoir de composantes en commun (dans ce cas, la distance qui les sépare sera de 1). Dans le cas où $\vec{A_{tr}}$ et $\vec{B_{tr}}$ ont au moins une composante en commun nous pouvons calculer deux différences :

La différence de rang : C'est l'écart de classement entre une composante précise dans $\vec{A_{tr}}$ et la même composante dans $\vec{B_{tr}}$.
i est le rang de C_t une composante de $\vec{A_{tr}}$ et $\rho(i)$ le rang de la même composante dans $\vec{B_{tr}}$, alors nous avons la différence de rang $E_{i,\rho(i)}$:

$$E_{i,\rho(i)} = \frac{(i - \rho(i))^2}{Nb^2 + (1 + \frac{i}{2})} \tag{4.14}$$

où Nb est le nombre de composantes conservées.

La différence d'intensité : Il nous faut également comparer la différence d'intensité des différentes composantes communes.
Pour cela nous considérons a_i l'intensité de la composante C_t de rang i dans $\vec{A_{tr}}$ et $b_{\rho(i)}$ l'intensité de la même composante dans $\vec{B_{tr}}$ (et dont le rang est $\rho(i)$), alors nous avons la différence d'intensité $I_{i,\rho(i)}$:

$$I_{i,\rho(i)} = \frac{\left|a_i - b_{\rho(i)}\right|}{Nb^2 + (\frac{1+i}{2})} \tag{4.15}$$

La concordance : La concordance $P(\vec{A_{tr}}, \vec{B_{tr}})$ est une première étape dans la mesure de la proximité entre $\vec{A_{tr}}$ et $\vec{B_{tr}}$. Elle est définie comme suit :

$$P(\vec{A_{tr}}, \vec{B_{tr}}) = (\frac{\sum_{i=0}^{Nb-1} \frac{1}{1+E_{i,\rho(i)} * I_{i,\rho(i)}}}{Nb})^2 \tag{4.16}$$

$P(\vec{A_{tr}}, \vec{B_{tr}})$ n'est pas symétrique.

La mesure de concordance : La concordance $P(\vec{A_{tr}}, \vec{B_{tr}})$ se concentre sur l'intensité et le rang des composantes et n'a pas la notion de direction que possède la distance angulaire. Nous introduisons donc la notion de direction en combinant la concordance avec la distance angulaire. Ainsi si $\delta(\vec{A}, \vec{B})$ est la distance angulaire entre $\vec{A}$ et $\vec{B}$, nous avons alors la mesure de concordance $\Delta(\vec{A_{tr}}, \vec{B_{tr}})$:

$$\Delta(\vec{A_{tr}}, \vec{B_{tr}}) = \frac{P(\vec{A_{tr}}, \vec{B_{tr}}) * \delta(\vec{A}, \vec{B})}{\beta * P(\vec{A_{tr}}, \vec{B_{tr}}) + (1-\beta) * \delta(\vec{A}, \vec{B})} \tag{4.17}$$

où β est un cœfficient permettant de donner à $P(\vec{A_{tr}}, \vec{B_{tr}})$ une importance plus ou moins grande en fonction des besoins.
$\Delta(\vec{A_{tr}}, \vec{B_{tr}})$ n'est pas symétrique.
$\Delta(\vec{A_{tr}}, \vec{B_{tr}})$ a été conçu au départ dans un contexte de classification, afin de comparer des vecteurs de textes à des vecteurs de classes. Dans un tel contexte, on cherche à trouver les similarités entre le vecteur de texte et le vecteur de classe. Ce dernier étant prépondérant. La symétrie n'est donc pas nécessaire et, par conséquent $\Delta(\vec{A_{tr}}, \vec{B_{tr}})$ est une mesure qui ne présente pas toutes les propriétés mathématiques d'une distance.
Dans notre contexte de segmentation de texte, un segment n'est pas plus important que celui qui le précède ou lui succède. Il est donc indispensable de disposer d'une mesure symétrique, et si possible disposant de toutes les propriétés d'une distance mathématique. Ce résultat peut s'obtenir facilement en définissant la distance de concordance comme suit :

$$D(\vec{A}, \vec{B}) = \frac{\Delta(\vec{A_{tr}}, \vec{B_{tr}}) + \Delta(\vec{B_{tr}}, \vec{A_{tr}})}{2} \tag{4.18}$$

On notera que $D(\vec{A}, \vec{B})$ est une application de type $d : E \times E \rightarrow [0, 1]$ comme la distance angulaire normalisée.

4.4.5.4 Paramètres et cas particulier de la distance de concordance dans Transeg

Lors de l'implémentation, il a fallu choisir les différents paramètres retenus pour la distance de concordance. Ces paramètres ont été choisis de manière empirique certes, mais pas au hasard.
Ainsi, pour le nombre de composantes conservées Nb, nous conservons $\frac{1}{3}$ de la taille initiale du vecteur. Nous avons constaté qu'en dehors des cas limites (tout conserver ou ne conserver presque rien), l'impact de la variation de Nb sur nos résultats était très minime (de l'ordre de quelques dixièmes de point). En étudiant les vecteurs sémantiques, nous avons constaté que, dans leur grande majorité, ils ne portaient d'informations significative que sur $\frac{1}{4}$ à $\frac{1}{2}$ de leur taille totale, le reste étant soit rempli de zéros, soit de valeurs extrêmement faibles. $\frac{1}{3}$ nous est apparu comme une valeur raisonnable, éliminant en moyenne le bruit tout en conservant aussi suffisamment d'informations.
Pour ce qui est du β, donc le poids de la mesure de concordance par rapport à la distance angulaire, nous avons choisi la valeur de $0,75$. Comme nous avons mené des expériences avec la distance angulaire comme distance thématique, nous voulions privilégier la notion de concordance dans la formule pour pouvoir comparer plus facilement les différences entre ces deux distances.

4.4.5.5 Environnement, automobile et distance de concordance

Reprenons notre exemple précédent sur l'automobile et l'environnement et comparons nos deux phrases avec la distance de concordance.
Les deux phrases n'ont pas changé, mais nous devons maintenant étendre un peu leur représentation.

- Ainsi, «*L'impact désastreux de l'automobile sur l'environnement, notamment sur le réchauffement climatique, n'est plus à démontrer.*» qui était initialement représentée par le vecteur $V_1 = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0]$ est maintenant représentée par une table de hachage prenant cette forme de :

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0

FIGURE 4.8 – Vecteur sémantique de la phrase 1

- De même, «*Notre nouveau modèle d'automobile ne rejette que* $127g$ *de* CO_2 *au kilomètre, faites un geste pour l'environnement, achetez une automobile X.*» qui

était initialement représentée par le vecteur $V_1 = [0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0]$ est maintenant représentée par une table de hachage prenant cette forme de :

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0

FIGURE 4.9 – Vecteur sémantique de la phrase 2

On conserve donc précieusement la position de chaque composante des vecteurs avant le tri et la réduction, pour pouvoir effectuer les comparaisons plus tard. Nous pouvons maintenant construire nos vecteurs trié-réduits pour obtenir ceci :

Déjà nous voyons, dès la représentation, les deux phrases se détacher légèrement l'une

10	5	1	2	3
2	1	0	0	0

FIGURE 4.10 – Vecteur sémantique de la phrase 1 trié et réduit

5	10	1	2	3
2	1	0	0	0

FIGURE 4.11 – Vecteur sémantique de la phrase 2 trié et réduit

de l'autre. Intégrer le rang des composantes dans la comparaison des deux vecteurs nous a permis d'intégrer une notion de hiérarchie au sein des concepts présents dans la phrase. Ainsi, le concept d'environnement subordonne celui d'automobile dans la première phrase, alors que c'est le contraire dans la deuxième. Si nous calculons la distance de concordance entre ces deux phrases nous obtenons une distance de concordance $D(\vec{Ph}_1, \vec{Ph}_2) \approx 0,56$. Si nous comparons cette distance avec la distance angulaire normalisée, cette dernière ayant les mêmes ensembles d'arrivée et de départ, nous remarquons que la distance de concordance est bien plus élevée que la distance angulaire. L'introduction de l'information portée par les rangs respectifs des différents concepts activés dans ces deux phrases a donc permis de les différencier avec plus de précision, et donc de retrouver la différence indéniable qui existe entre les deux.

La distance de concordance nous permet donc de différentier deux phrases sémantiquement proches, mais thématiquement distinctes. Elle est donc particulièrement adaptée à la segmentation thématique.

4.5 Le développement de Transeg

Transeg a été développé en JAVA, sous l'environnement de développement Eclipse. Plusieurs raisons ont motivé ce choix :

- **La portabilité.** Le principal avantage de JAVA est d'être un langage portable, sans qu'aucune modification ne soit nécessaire, sur la grande majorité des systèmes utilisés à l'heure actuelle (Linux, Unix, Windows, MacOS, etc.). Cela nous a permis de bénéficier d'un grande liberté de développement et changer régulièrement de plate forme sans difficulté. Nous avons pu ainsi développer sous un environnement Windows sur un ordinateur portable modérément puissant et effectuer nos expériences sous un environnement Linux sur un serveur dédié. Eclipse étant un logiciel développé sous JAVA, il est présent sur toutes ces plate formes, nous permettant d'en changer sans changer d'environnement.
- **La facilité de développement.** JAVA dispose de nombreuses librairies de classes qui implémentent nombre d'outils utilisés dans le TALN. La librairie WEKA ([Garner, 1995]), que nous décrirons plus en détail plus loin dans cette section, notamment nous aura été très utile. Plus généralement, JAVA offre de nombreux outils, allant de la gestion des communications via un réseau à l'usage des expressions régulières, qui nous ont permis de gagner du temps dans le développement. L'environnement de développement Eclipse, quant à lui, offre un confort de développement appréciable lors des longues séances de « codage ».
- **Le caractère libre.** Eclipse est un logiciel sous licence libre[37] ce qui nous libère des contraintes de droits, de licences et autre tracasseries juridiques que des environnements propriétaires auraient pu nous causer.

Dans cette section, nous présentons certains aspects techniques du développement de Transeg et les motivations de certains de nos choix.

4.5.1 Interface avec SYGFRAN

SYGFRAN, que nous avons décrit dans la section 4.3, est l'application qui nous fournit les vecteurs sémantiques pour les phrases analysées. Il nous était donc nécessaire de connecter les deux applications (SYGFRAN et Transeg) entre elles pour arriver au résultat escompté.
Il nous était possible d'intégrer SYGFRAN directement à notre application comme un module. En effet, bien que SYGFRAN soit développé en langage C, il est possible d'intégrer des modules en langage C dans une application JAVA. Toutefois cela nous aurait

37. EPL ou Eclipse Public License

fait perdre tout l'aspect portable de l'application Transeg. De plus, Transeg doit pouvoir fonctionner avec d'autres représentations vectorielles de la phrase. Il nous fallait donc séparer l'application Transeg de l'application qui lui fournit sa représentation vectorielle de la phrase. Pour cela nous avons opté pour une interface entre Transeg et son « fournisseur de vecteurs » (que ce soit SYGFRAN ou une autre application) qui passe par le réseau.

Transeg transmet donc une à une les phrases du texte à SYGFRAN par socket et récupère les vecteurs sémantiques au fur et à mesure. Ces vecteurs sont sauvegardés dans un fichier sur la machine qui exécute Transeg, ce qui permet de n'avoir à effectuer l'analyse coûteuse de SYGFRAN qu'une seule fois par texte et ainsi de pouvoir tester plusieurs configurations différentes sans devoir repasser par une phase d'analyse.

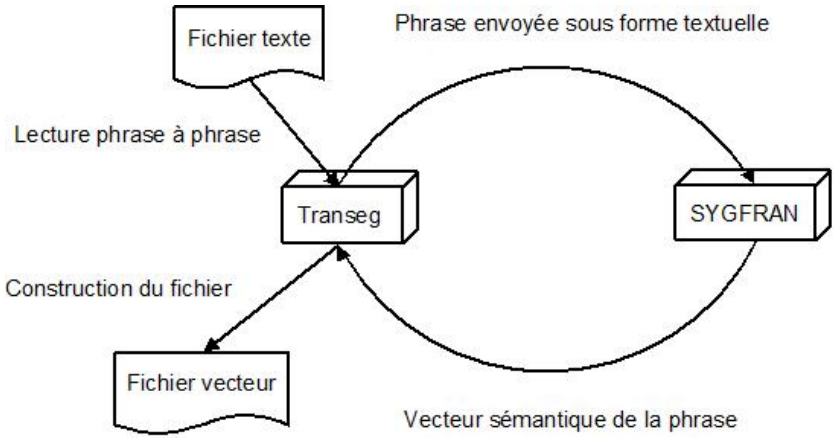

FIGURE 4.12 – Echange entre Transeg et SYGFRAN

4.5.2 Implémentation des algorithmes

Le cœur de l'application est le système de segmentation de texte en lui-même. Comme nous suivons assez fidèlement les algorithmes présentés dans le chapitre précédent, nous ne présenterons ici que certaines spécificités techniques ou certains ajustements nécessaires au bon fonctionnement du système.

4.5.2.1 Détection de changement de thème par identification des zones de transition : un algorithme par glissement de fenêtre

Le principal problème qu'il nous a fallu gérer lors de l'implémentation de cet algorithme fut comme souvent la gestion des effets de bords.
Comme notre algorithme se base sur une fenêtre qui « glisse » le long du texte et qui travaille sur le centre de cette fenêtre, les extrémités du texte posent problème. Nous avons

donc décidé de toujours considérer la première phrase du texte comme une frontière thématique (c'est la frontière du premier segment thématique). De même, nous considérons que la dernière phrase du texte ne peut jamais être un segment thématique. Ce faisant le produit du score de transition de la première phrase du texte par le score de rupture de la phrase qui la précède (qui par définition n'existe pas) est toujours 1 et ce même produit pour la dernière phrase est toujours 0.
Ces ajustements particuliers à la première et à la dernière phrase du texte ne sont toutefois pas suffisants. Comme notre algorithme s'intéresse au centre de la fenêtre et compare les deux moitiés de celle-ci, il est inévitable que ces deux moitiés soient de tailles différentes aux extrémités du texte. Pour ne pas favoriser une moitié de la fenêtre par rapport à une autre nous avons décidé de ne jamais considérer dans la plus grande moitié plus de phrases qu'il n'y en a dans la plus petite. Ainsi, seules les p premières phrases de la plus grande moitié de fenêtre, où p est le nombre de phrases dans la plus petite moitié (figure 4.13) sont considérées.
Cette astuce est rendue possible par le postulat sur l'organisation d'un segment théma-

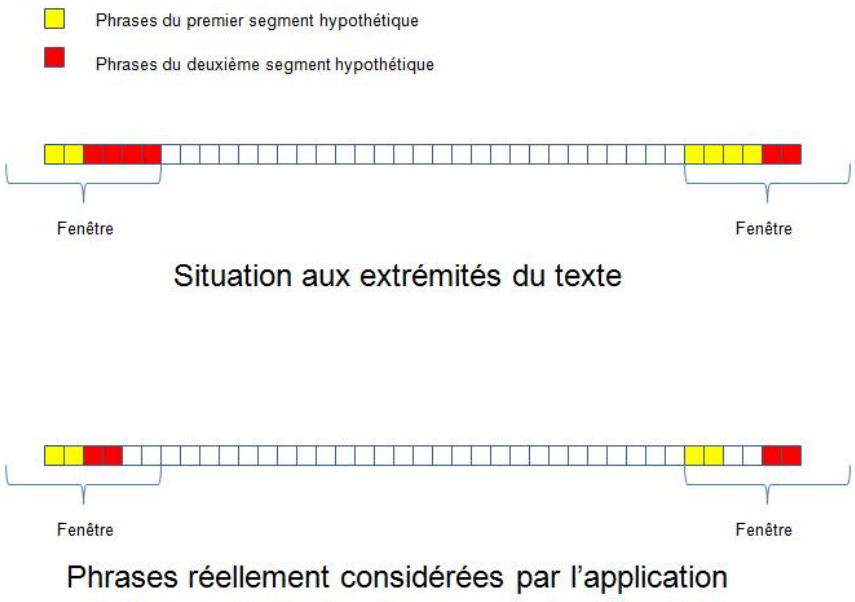

FIGURE 4.13 – Gestion des effets de bord

tique du chapitre 3. En effet, comme les premières phrases d'un segment sont les plus importantes, il est possible de laisser de coté certaines des dernières phrases sans perdre trop d'information.

4.5.2.2 WEKA et structure des classes de segmentation

Dans la mesure où les algorithmes de clustering (stricts ou flous) que nous utilisons sont couramment utilisés, nous avons décidé d'utiliser la bibliothèque JAVA Weka

([Garner, 1995] `http://www.cs.waikato.ac.nz/ml/weka/`) pour tester leur efficacité dans le cadre de notre tâche.

Weka est une collection d'algorithmes de fouille de données. On y trouve des algorithmes de clustering (ceux que nous utilisons notamment), mais aussi des algorithmes de visualisation, de pré-traitement des données, de classification ou encore de règles d'association. Weka présente l'avantage de pouvoir être directement importée dans un programme JAVA ou d'être utilisée indépendamment. Nous avons fait le choix d'importer directement les classes que nous utilisons pour les intégrer à notre application.

Les algorithmes de clustering de Weka sont des classes JAVA héritant toutes d'une classe

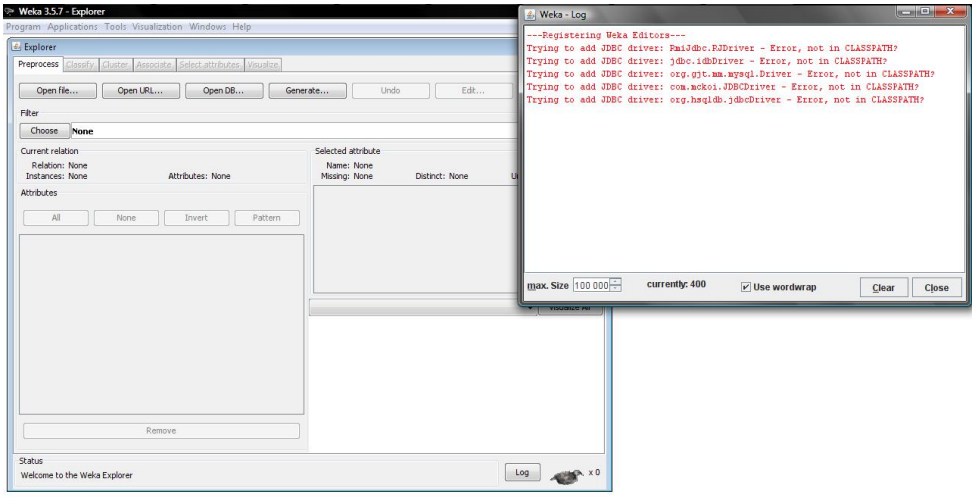

FIGURE 4.14 – L'interface de Weka

mère. Ainsi, tous les algorithmes utilisent les mêmes données en entrée et restituent les mêmes données en sortie. Nous avons respecté cet esprit de standardisation dans le développement de notre application, et donc nous n'avons pas une classe de segmentation thématique qui utilise différents algorithmes, mais plusieurs classes qui héritent en cascade des propriétés d'une classe abstraite « Segmenteur thématique »(figure 4.15).

La construction des résultats se fait en suivant la stratégie présentée dans le chapitre 3.

4.5.2.3 Fusionner les approches

Les approches par détection active des frontières (notre approche par calcul de distance) et passive des frontières (celles utilisant le clustering) étant complémentaires, il nous est paru pertinent de tenter de fusionner ces deux visions. Toutefois, comme le clustering strict et le clustering flou donnent des résultats différents, nous avons mis en place

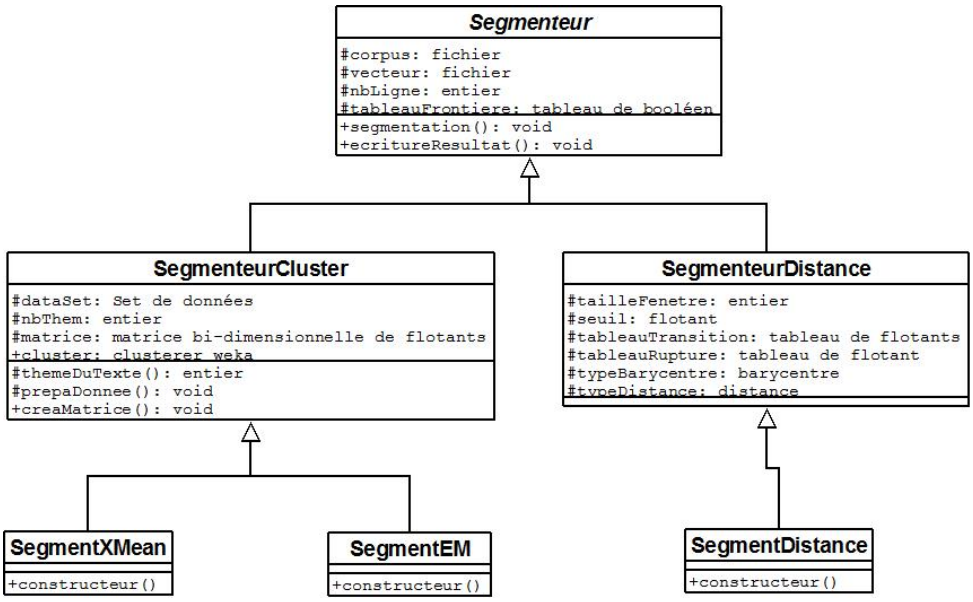

FIGURE 4.15 – Schéma UML des principales classes de segmentation thématique

des stratégies différentes.
Ce paragraphe présente les stratégies de fusion de notre approche par calcul de distance avec des algorithmes de clustering strict et flou. Nous présentons également une amélioration possible de notre approche par calcul de distance s'inspirant des techniques de classification hiérarchique.

« Fusion » avec un algorithme strict : la stratégie de la validation

L'algorithme de clustering strict nous sert ici d'outil de validation. Nous comparons les frontières trouvées par l'algorithme par calcul de distance à celles trouvées par l'algorithme de clustering et nous « validons » les frontières issues du calcul de distance si elles se trouvent à moins de deux phrases d'une frontière ramenée par le clustering.
Nous avons choisi de faire valider les frontières issues de la détection active par la détection passive, plutôt que le contraire, car la détection active s'intéresse spécifiquement aux propriétés des frontières et doit donc avoir plus de chance de trouver la bonne position.
Le choix de deux phrases de marge est une conséquence du choix d'une de nos mesures d'évaluation, le FScore flou, présenté dans le chapitre prochain, qui considère les phrases autour de la frontière attendue plutôt que juste la frontière elle-même.
Cette fusion s'apparente donc plus à un système de vote qu'à une réelle hybridation des méthodes.

Combiner le clustering flou et les scores de transition / rupture

Le clustering flou nous donnant des scores de probabilités, combiner les résultats de ce dernier avec ceux de notre algorithme par calcul de distance. En effet, même si le produit *score de transition d'une phrase i* × *score de rupture de la phrase précédente* n'est pas une probabilité, il représente tout de même la « chance » que l'on change de thème, on le notera C_i. On peut donc considérer que $1 -$ *score de transition d'une phrase i* × *score derupture de la phrase précédente* comme un score représentant lui la « chance » que l'on ne change pas de thème, que l'on notera $1 - C_i$.

Dés lors, en considérant la matrice des probabilités d'appartenance des phrases du texte à un thème comme les poids des sommets d'un graphe orienté et les scores C_i et $1 - C_i$ comme les valeurs des arcs de ce graphe, nous pouvons chercher un plus long chemin dans ce graphe pour en quelque sorte « cartographier » les thèmes du texte (figure 4.16).

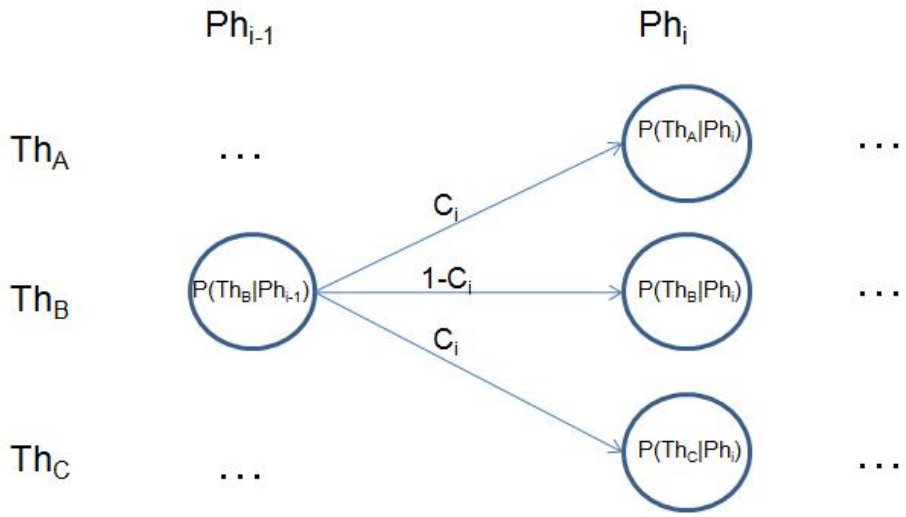

FIGURE 4.16 – Extrait du graphe orienté

Amélioration par classification hiérarchique

Bien que n'étant pas, à proprement parler, une fusion entre un algorithme de clustering et notre approche linéaire par calcul de distance, nous avons tenté d'améliorer les résultats de cette dernière en effectuant un post-traitement de type classification hiérarchique [38].

Le principe, relativement simple, consiste à considérer une frontière identifiée par notre approche linéaire et les deux segments qu'elle sépare. On cherche alors dans le texte pour chacun des deux segments observés, le segment qui en est thématiquement le plus proche. On fusionne ces paires de segments thématiquement proches pour créer deux pseudo-

38. La classification hiérarchique étant une méthode parfois utilisée dans les tâches de clustering, présenter cette variante de notre approche ici semblait judicieux.

modèles de thèmes. Cette fusion n'étant qu'un calcul de barycentre entre les vecteurs sémantiques des segments concernés.
En considérant comme des modèles de thèmes les deux nouveaux vecteurs sémantiques, nous les comparons avec chacune des phrases de la zone de transition. Leur rapprochement / éloignement avec chacun des modèles nous permettra d'ajuster la place de la frontière dans la zone de transition.
Cette variante de notre approche peut être perçu comme une amélioration du score de rupture qui ne considérerait plus le texte de manière linéaire.

4.5.3 Interface graphique de Transeg

Afin de faciliter l'usage de Transeg, nous avons implémenté une interface graphique pour l'application. Cette interface graphique suit la même philosophie que l'interface réseau de Transeg avec SYGFRAN : elle se veut modulaire et évolutive.
Elle est donc totalement indépendante de la partie du programme qui effectue la tâche de segmentation et, de fait, propose différentes options de segmentation thématique. Il est par exemple possible de choisir les différents paramètres de Transeg (taille de la fenêtre, seuil, type de distance utilisée, etc.), de sélectionner un algorithme autre que Transeg (clustering, clustering flou, hybridation, etc.). Un outil d'évaluation des résultats est également disponible[39].
En plus de permettre d'ajuster tout les paramètres proposés par notre approche (figure

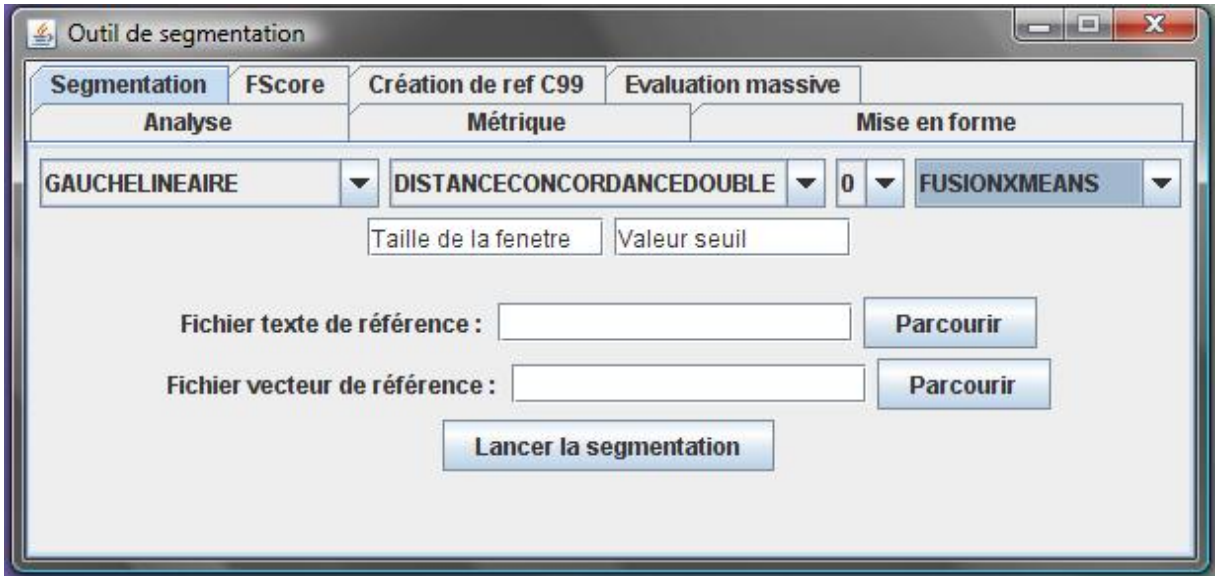

FIGURE 4.17 – L'interface graphique de Transeg

4.17), l'interface permet d'accéder rapidement à un système d'évaluation des résultats ou

39. A condition, évidemment, de disposer des frontières de référence pour le texte.

encore de mise en forme du texte.

4.6 Conclusion

Dans ce chapitre nous avons présenté la mise en pratique du cadre théorique exposé dans le chapitre 3.
Dans la section 4.2, nous avons fait une présentation globale de l'architecture de l'application développée pour cette mise en pratique. Cette architecture en trois phases découle directement de la manière dont nous avons découpé le problème que nous nous sommes posés.
Dans la section 4.3 nous avons présenté la manière dont les vecteurs sémantiques sont générés par l'application SYGFRAN. SYGFRAN est un analyseur morpho-syntaxique de la phrase française qui se base sur des règles proches de la grammaire française et le système opérationnel SYGMART pour présenter le résultat de cette analyse sous forme d'arbre. Cette analyse morpho-syntaxique sous forme d'arbre permet ensuite de construire les vecteurs sémantiques en combinant linéairement et successivement des vecteurs sémantiques de mots puis de groupes. Les vecteurs sémantiques des mots sont obtenus à partir d'une projection de l'espace lexical dans un espace de concepts défini par un thésaurus.
En section 4.4, nous avons montré que les outils de comparaison de vecteurs que nous avions à notre disposition n'étaient pas forcément adaptés à notre tâche. Nous y présentons donc un outil de comparaison plus pertinent dans une tâche de segmentation thématique : la distance de concordance. Cette distance est plus discriminante que les mesures classiques grâce à l'intégration de la notion de rang relatif des composantes des vecteurs dans le calcul de leur distance.
Dans la dernière section de ce chapitre, la section 4.5, nous avons offert une vue d'ensemble des différents éléments notables propres au développement de l'application.
Dans le chapitre suivant nous allons utiliser les outils présentés ici pour tester notre cadre théorique sur un corpus de texte.

5

De l'évaluation automatique au jugement humain

Sommaire

5.1 Introduction

L'évaluation d'une tâche de TALN est toujours un défi en soi. Le langage naturel est un phénomène généré par l'humain à destination de l'humain, son interprétation est subjective. L'évaluation de méthodes automatiques destinées à l'analyser, le manipuler ou le générer doit donc, dans la mesure du possible, prendre en compte cette subjectivité. C'est là, que se situe toute la gageure de la tâche d'évaluation, savoir rester objectif et rigoureux, tout en ne perdant pas de vue que le phénomène étudié n'est pas objectivement évaluable.

Nous consacrerons la section 5.2 aux différents outils à notre disposition pour évaluer les résultats d'un travail de segmentation thématique. Nous y aborderons les problèmes relatifs à la création d'un corpus de référence et également les différentes mesures utilisées dans l'évaluation de la segmentation thématique.

Dans la section 5.3, nous présenterons le corpus sur lequel nous avons mené nos expériences. Ce corpus est en réalité composé de textes de deux origines différentes tant au

niveau du style que dans la manière dont les références de ces textes ont été construites. Dans la section 5.4, nous présenterons ce qui a motivé notre choix de c99 comme baseline, ainsi que les différents ajustements que nous avons fait sur le corpus pour que la comparaison soit la plus loyale possible.
La section 5.5 présentera les résultats de notre approche et de ses différentes variantes. Nous y ferons varier de nombreux paramètres afin de voir leur incidence sur les résultats et ainsi justifier certains de nos choix.
Nous consacrerons la section 5.6 à présenter notre tentative de mettre en place une évaluation humaine. Nous insisterons notamment sur le protocole imaginé pour l'occasion et sa justification, mais nous verrons également la manière dont nous l'avons mis en œuvre et les résultats mitigés que nous avons obtenus.
Nous conclurons ce chapitre dans la section 5.7.

5.2 Évaluer la segmentation thématique de texte

Pour évaluer correctement la tâche de segmentation thématique, il est nécessaire de disposer d'un corpus de textes pré-segmentés qui servira de référence et d'une mesure d'évaluation de la qualité du résultat vis à vis de cette référence. Ces deux aspects de l'évaluation influent grandement sur les résultats, il faut donc les choisir avec beaucoup de soin.
Cette section présentera donc les différents outils que nous avions à notre disposition pour la mise en place de notre évaluation.

5.2.1 La création de corpus de référence

La création d'un corpus pour tester une méthode de TALN est toujours un soucis majeur. En effet, pour évaluer la segmentation thématique nous avons besoin de textes qui ont été pré-segmentés. Ce type de ressources est très difficile à trouver, car très coûteuse à produire en terme d'investissement humain. C'est d'autant plus vrai en français (notre langue d'application), qui ne disposant pas d'une visibilité internationale comme l'anglais, ce voit attribuer moins de moyens notamment pour produire ce genre de ressources.
Nous présenteront donc ici une alternative à la création manuelle d'un corpus de référence ainsi que deux approches de la création de corpus de référence par l'homme.

5.2.1.1 La concaténation de textes courts

La solution la plus courante pour palier le manque de corpus de référence en segmentation thématique est la création de corpus que nous nommerons « artificiels ». Ces corpus sont composés en concaténant plusieurs textes courts traitant de sujets différents. On considère alors chaque texte comme un segment thématique.
Cette méthode de création d'un corpus de référence présente le double avantage d'être simple et peu coûteuse à mettre en œuvre. Il est aisé de récupérer sur des sites Internet d'information, par exemple, un grand nombre d'articles de petite taille traitant de sujets divers. On peut disposer dès lors de nombreux « segments thématiques » les uns à la suite des autres et parfaitement identifiés, sans erreur ni contestation possible. Cette méthode est très utilisée dans la littérature ([Choi, 2000]).
Le problème de cette méthode, c'est qu'elle ne fournit pas un cadre de test pour une tâche de segmentation **thématique** de textes, mais pour de la segmentation de textes. En effet, les textes venant d'auteurs différents et n'ayant (dans la majorité des cas) rien à voir les uns avec les autres, retrouver ces textes ne peut raisonnablement être considéré comme une recherche de la structure thématique d'un texte, mais plutôt comme une tâche de segmentation d'un flux de données, ou de recherche d'auteur.
Ce type de corpus est donc peu satisfaisant dans le cadre de notre tâche. Nous cherchons à trouver quels sont les indices au sein d'un texte qui nous permettraient de retrouver les thèmes abordés dans celui-ci. Or, un texte est une construction bien plus complexe qu'une simple concaténation d'idées.

5.2.1.2 La référence de l'expert

Lorsque l'on veut disposer de textes pré-segmentés pour servir de référence dans une expérience de TALN, faire appel à un expert pour effectuer cette segmentation humaine parait logique. Cet expert peut être un linguiste, mais aussi l'auteur même du texte. Nous avons en effet décidé de considérer l'auteur d'un texte comme un expert. Dans la mesure où ce dernier a une vision claire du message qu'il souhaite faire passer dans son texte, il peut être considéré comme expert au moins lorsqu'il s'agit d'analyser, résumer ou segmenter thématiquement ses propres productions.
C'est donc l'expert désigné qui identifie les segments thématiques en se basant sur des critères linguistiques précis, dans le cas d'un appel à un linguiste ou, si l'on se base sur l'expertise de l'auteur du texte, sur le message que ce dernier veut communiquer.
Ce type de corpus présente l'avantage de nous proposer des références faciles à justifier devant la communauté et donc nous permet d'évaluer les résultats de nos expériences de manière fiable.

Toutefois, la faille de tels corpus réside dans le caractère très subjectif de la tâche que nous avons à évaluer. Si dans certains domaines l'avis de l'expert est difficilement contestable, ce n'est pas le cas de la segmentation thématique. Comme souvent dans ce type de tâche, si l'on demande à plusieurs experts la même analyse, on aura probablement autant de résultats différents que d'experts. Peut être pas très éloignés les uns des autres, mais différents tout de même. Le problème se pose alors de savoir lequel des résultats choisir. Doit-on n'en prendre qu'un et considérer les autres comme faux ? Sont-ils tous justes et dans ce cas comment évaluer les résultats d'une méthode automatique sur la base de ces références ?
Comme nous le verrons plus loin, nos textes segmentés par des experts, l'ont été par leurs auteurs. Nous n'avons donc qu'une seule opinion, nous n'avons pas à choisir entre plusieurs expertises. Toutefois, cela pose la question de savoir si un résultat non conforme avec l'avis de l'expert est vraiment mauvais, et si c'est le cas dans quelle proportion. C'est malheureusement une des contraintes de notre tâche que nous sommes obligés de prendre en compte dans notre évaluation.

5.2.1.3 La référence consensuelle

Une autre possibilité serait de s'appuyer sur des textes segmentés par « consensus ». Chaque texte serait ainsi présenté à un panel de juges qui le segmenterait individuellement. On regrouperait alors les résultats et ne considérerait comme des frontières acceptables que celles qui auraient été désignées par la majorité des juges. C'est ainsi que procède [Bestgen & Piérard, 2006] pour créer son corpus de test des différents algorithmes de segmentation thématique par exemple.
L'avantage d'un corpus consensuel, c'est qu'il représente une « moyenne » de ce que l'utilisateur moyen s'attend à recevoir comme solution.

5.2.2 « Mesurer » les résultats

Le deuxième aspect important dans l'évaluation d'une tâche comme la segmentation thématique est sans doute la mesure utilisée. En effet, nous l'avons déjà évoqué à plusieurs reprises, la détermination des segments thématiques d'un texte est une tâche très subjective. Plusieurs solutions sont donc en général possibles. Toutefois, nous n'aurions pas envisagé de traiter un tel problème si nous ne pensions pas possible d'automatiser le processus, ce qui sous entend qu'il existe tout de même une solution de référence (voir section précédente). Pour prendre en compte la subjectivité de la tâche tout en utilisant une référence unique, nous devons nous doter de mesures de qualité qui ne considèrent

pas systématiquement comme faux un simple décalage d'une ou deux phrases entre la frontière proposée en solution et celle de référence.
Nous présentons donc ici deux mesures qui visent à intégrer cet aspect flou nécessaire à l'évaluation des résultats proposés par une application de segmentation thématique.

5.2.2.1 WindowDiff

WindowDiff est une évolution de la mesure P_k proposée par [Beeferman *et al.*, 1997]. La mesure P_k se propose de de prendre en compte la distance entre une frontière ramenée par l'algorithme et la frontière que ce dernier aurait du ramener. Pour cela, elle évalue la probabilité d'erreur sur la segmentation en considérant la probabilité que deux phrases éloignée d'une distance k [40] d'être dans un même segment du document de référence (ref) et du document résultat (hyp). La formule pour un corpus de longueur n étant la suivante :

$$P_k(ref, hyp) = \sum_{1 \leq i \leq j \leq n} D_k(i,j)\delta_{ref}(i,j)\overline{\oplus}\delta_{hyp}(i,j) \tag{5.1}$$

Dans cette formule $\delta_{ref}(i,j)$ et $\delta_{hyp}(i,j)$ sont des fonctions prenant la valeur 1 si les deux indice i et j appartiennent au même segment dans leurs corpus respectifs (ref et hyp) et 0 si ce n'est pas le cas. L'opérateur $\overline{\oplus}$ est l'opérateur **XNOR** qui suit la table 5.1. La

1	XNOR	1	=1
1	XNOR	0	=0
0	XNOR	1	=0
0	XNOR	0	=1

TABLE 5.1 – L'opérateur XNOR : les deux ou aucun des deux

fonction D_k est une distribution de probabilité de distance entre phrases sur un ensemble de phrases choisi aléatoirement sur le corpus de référence (ref). Elle dépend du paramètre k qui est à la taille moyenne d'un segment (toujours sur le corpus de référence ref). Si D_k est uniforme le long du texte, alors elle correspond à la probabilité que deux phrases prises aléatoirement dans le texte soient dans deux segments différents (toujours en fonction de k et de leur distance).
La mesure P_k présente plusieurs failles mises en évidence par [Pevzner & Hearst, 2002]. Parmi ces failles, on notera le fait qu'elle est peu claire, qu'elle comptabilise différemment des erreurs de même type ou encore que les ajouts de très petits segments ne sont pas ou peu comptabilisés. Aussi, proposent-ils une mesure appelée WindowDiff qui est inspirée de la mesure P_k, mais qui pallie certain de ses défauts.

40. Cette distance est un nombre de phrases dans notre cas, mais peut également être un nombre de mots, de paragraphes ou de tout autre unité en fonction du type de segmentation.

WindowDiff est une mesure beaucoup plus simple qui utilise une fenêtre de longueur k glissant le long des deux textes et comptabilisant à chaque décalage d'une phrase la différence du nombre de frontières entre le texte de référence (ref) et celui produit par la méthode (hyp) :

$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})|) \tag{5.2}$$

où N est le nombre de phrases du texte et $b(x_i, x_j)$ une fonction donnant le nombre de frontières du texte x entre les phrases i et j.
Le problème de cette mesure est qu'elle peut être supérieure à 1 et donc qu'elle ne représente pas un taux d'erreur et ne permet pas de comparer efficacement les résultats sur des textes de tailles très différentes. Nous avons donc utilisé une variante de cette mesure qui, plutôt que de considérer l'écart entre le nombre de frontières à chaque décalage de frontière comptabilise une erreur lorsque cet écart est non nul et aucune dans le cas contraire. La nouvelle formule est donc :

$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (O(b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}))) \tag{5.3}$$

où $O(x)$ est une fonction prenant la valeur 0 lorsque x est nul et 1 dans les autres cas.
Le principal problème de la mesure WindowDiff est qu'elle comptabilise plusieurs fois la même erreur. En effet, si par exemple un algorithme de segmentation thématique ramène comme phrase frontière une phrase qui se trouve être adjacente à la phrase considérée comme référence, cette erreur sera comptabilisée deux fois, quatre fois s'il y a un écart d'une phrase. Or, l'objectif de cette mesure est justement d'offrir une certaine souplesse dans l'évaluation.
Un autre problème de WindowDiff, celui-là plus dans l'interprétation du résultat que dans la qualité de ce dernier, est l'incapacité avec cette mesure à évaluer où ce situe la défaillance d'une méthode étudiée. WindowDiff comptabilise indifféremment les erreurs relatives au bruit et celles provenant du silence. Ainsi, si cette mesure permet d'avoir un indice de qualité final sur les résultats d'une méthode, elle est inappropriée au cours du développement.
Dans nos expériences nous utiliserons la mesure WindowDiff avec $k = 5$. C'est la valeur couramment utilisée dans les diverses évaluations d'algorithmes de segmentation thématique présentes dans la littérature.

Comme beaucoup de mesures d'évaluation elle n'a que peu de valeur seule et doit être combinée avec d'autres mesures pour pouvoir bénéficier d'une vision plus précise de la

qualité des résultats. Nous avons donc décidé d'évaluer nos résultats en prenant aussi en compte une autre mesure comme par exemple le F_{Score}.

5.2.2.2 Rappel, précision et F_{Score} : variantes adaptées à la segmentation thématique

Le F_{Score} (ou FMesure) est une mesure couramment utilisée dans les évaluations en TALN. C'est une mesure qui peut être vue comme un compromis entre les deux grands standards de l'évaluation que sont le rappel et la précision ([BY1999]).
Dans la majorité des cas, lorsque l'on favorise le rappel, la précision en pâtit et inversement. Le F_{Score}, en étant une combinaison des deux, permet d'évaluer la performance globale d'une méthode sur ces deux mesures. Ainsi, on peut évaluer si le gain dans une des mesures et valable vis à vis de la perte dans l'autre.
Dans le cadre de la segmentation thématique la précision s'exprimerait donc ainsi :

$$Précision = \frac{Nombre\ de\ frontières\ justes\ ramenées}{Nombre\ de\ frontières\ ramenées} \tag{5.4}$$

De même que l'on écrirait le rappel de la sorte :

$$Rappel = \frac{Nombre\ de\ frontières\ justes\ ramenées}{Nombre\ de\ frontières\ attendues} \tag{5.5}$$

Le F_{Score} s'écrivant donc :

$$\mathrm{F}_{Score}=(^2+1)\times\ Précision\ \times\ Rappel\frac{}{^2\times Précision + Rappel} \tag{5.6}$$

avec un paramètre permettant de donner plus de poids à la précision ou au rappel dans le cas d'évaluation cherchant à favoriser l'un ou l'autre. Nous n'utiliserons pas cette formule du F_{Score} dans notre évaluation, cette dernière étant inadaptée à la tâche, mais une variante que nous allons présenter par la suite. Toutefois, nous considérons comme d'une importance équivalente le rappel et la précision aussi le paramètre sera t'il à 1 lors de nos évaluations.

Le problème avec le F_{Score} dans une tâche telle que la segmentation thématique c'est que les frontières ramenées doivent correspondre exactement à celles du corpus de référence pour être comptabilisées comme juste. Aussi, les valeurs de rappel, de précision et donc de F_{Score} en segmentation thématique sont rarement satisfaisantes et peu, voir pas du tout, exploitables ([Beeferman *et al.*, 1999]). En effet, si une frontière identifiée par l'algorithme se trouve être voisine de la frontière de référence elle est tout simplement comptabilisée fausse.

Tout comme la mesure WindowDiff utilise une fenêtre pour laisser une marge de tolérance à l'erreur, le F_{Score} (et à travers lui le rappel et la précision) peut être modifié pour avoir une certaine tolérance à l'erreur. Ainsi, lors de l'évaluation DEFT'06 [41] ([Azé *et al.*, 2006]) les organisateurs, conscients des limites d'une application stricte du F_{Score}, ont proposé une variante du F_{Score} qui est tolérante vis à vis des erreurs marginales. Cette variante « floue » du F_{Score}, appelée F_{Score} **souple**, considère comme juste les frontières ramenées par l'algorithme évalué si celles-ci se trouvent dans une fenêtre autour de la frontière de référence. Dans le cadre de nos expériences nous avons évalué le F_{Score} souple pour une fenêtre contenant deux phrases avant et après la frontière à trouver. [Azé *et al.*, 2006] ont prouvé que si F_{Score} souple améliorait sensiblement les résultats, il ne changeait pas le classement des différentes méthodes évaluées lors de DEFT'06.
La grande force du F_{Score} souple par rapport à WindowDiff est qu'il est basé sur les notions éprouvées de précision et de rappel, même si ces dernières ont été assouplies. Il nous est donc possible d'évaluer les performances d'une méthode pas seulement en terme de taux d'erreur, mais aussi en terme de bruit et de silence. C'est donc une mesure plus adaptée lorsque l'on souhaite étudier en détail l'effet de la variation de certains paramètres sur une méthode. De plus le F_{Score} souple ne comptabilise pas plusieurs fois la même erreur.

Nous avons fait le choix d'utiliser les deux mesures d'évaluation que nous venons de présenter dans nos expériences. Cela nous permettra de mettre en évidence l'impact de la mesure d'évaluation sur le résultat.

5.3 Présentation du corpus d'expérimentation

Le choix d'un corpus d'expérimentation dans le TALN est crucial, c'est encore plus valable lors de l'évaluation d'une tâche de segmentation thématique. Ce corpus doit réunir plusieurs qualités indispensables afin que l'expérience puisse servir à valider les théories élaborées en amont.

- La taille du corpus doit être suffisante pour que les résultats obtenus aient une réelle signification. Des résultats, qu'ils soient bons ou mauvais, obtenus sur un corpus trop petit n'ont aucune valeur.
- Le corpus se doit d'offrir un minimum de variété dans le style et les auteurs des textes proposés. Un corpus trop uniforme risque d'entraîner un effet d'adaptation de la méthode au style du texte.

41. Pour DEfi Fouille de Texte, l'édition 2006 était autour de la segmentation thématique de documents.

– Le corpus doit être représentatif de la réalité à laquelle sera confrontée la méthode ou l'algorithme évalué. Le choix d'un corpus trop artificiel aura pour conséquence des résultats sans réel rapport avec l'efficacité véritable de la méthode ou de l'algorithme évalué.

A ces trois grandes contraintes de qualité que nous venons de définir il nous faut rajouter des contraintes liées à la disponibilité du corpus. Comme nous l'avons évoqué plus haut, les textes pré-segmentés pouvant servir de références sont rares et coûteux en ressources à produire, donc difficile à se procurer. Mais il faut aussi prendre en considération les soucis légaux qui peuvent intervenir lors de la constitution du corpus. Ces différents problèmes rendent compliquée la constitution d'un corpus d'évaluation crédible.
Le corpus sur lequel nous avons évalué notre méthode est en réalité composé de deux corpus que nous allons présenter ici.

5.3.1 Un corpus de textes politiques

Ayant participé à DEFT'06 ([Azé *et al.*, 2006]) nous avions accès à un vaste corpus de textes « segmentés thématiquement ». Le corpus de DEFT'06 est composé de trois grands ensembles de texte :

– Un corpus dit « juridique » composé de textes de loi. Les « segments thématiques » de ce corpus étant les lois, celui-ci est assimilable à un corpus artificiel obtenu par concaténation de textes. De plus chaque loi commençant par la mention *Article*[42], il est extrêmement facile d'obtenir un rappel maximum en ramenant toutes les phrases commençant par cette mention. Ce corpus était trop artificiel et trop éloigné de la tâche pour que nous l'utilisions ici. La segmentation de textes juridiques est une tâche plus adaptée à des méthodes spécialisées et en général supervisées.
– Un corpus dit « scientifique », qui se trouve être une thèse dont les paragraphes / sections sont considérés comme des segments thématiques. En dehors du fait que nous avons contesté la typographie comme unique indice de segmentation thématique dans le chapitre 3, ce texte est protégé par un copyright et il nous est interdit d'en présenter tout ou partie dans ce document. L'impossibilité de présenter ne serait-ce qu'un exemple, rendant peu crédible d'éventuels résultats présentés ici, nous avons décidé de ne pas l'intégrer dans notre évaluation.
– Un corpus dit « politique », qui rassemble un ensemble de discours d'hommes politiques segmentés par leur auteur. C'est dans ce corpus que nous avons décidé de

42. Une loi est composée de plusieurs articles, ces derniers sont normalement numérotés, mais pour des raisons évidentes les évaluateurs ont enlevé les numéros.

choisir les textes de notre expérimentation. Toutefois, il nous a fallu régler certains problèmes propres à la forme sous laquelle ce corpus nous a été proposé par les organisateurs de DEFT'06.

Extraction de 22 discours politiques

Le corpus politique de DEFT'06 est énorme, il se compose de 303 373 phrases et 7 199 480 mots, ce qui aurait pu constituer un excellent corpus pour nos évaluations. Toutefois, deux problèmes majeurs se posent lorsque l'on observe ce corpus tel qu'il nous a été donné :

- Tout d'abord, les discours sont concaténés, sans indice d'où chacun commence ou fini. Les organisateurs n'ont pas fait de différence entre les frontières à l'intérieur des textes et celles qui séparent les textes. Nous souhaitons évaluer une méthode qui segmente thématiquement **un seul** texte en se basant sur des postulats sur sa structure thématique et pas une méthode retrouvant la frontière entre deux textes différents. Cette concaténation n'est donc pas appropriée pour notre expérience.
- Le corpus qui nous a été fourni est extrêmement bruité. De nombreuses phrases sont entièrement en lettres capitales, certaines ne comportent que des signes de ponctuation et même certains passages n'ont aucun sens. Ce bruit, très présent (trop présent même), est probablement du à des erreurs d'éditions lors de la création du corpus. Il est aussi extrêmement handicapant dans le cadre de nos expériences. Certes, un corpus d'expérience en TALN doit comporter du bruit, puisque qu'il est rare que lorsque les méthodes développées dans cette discipline sont appliquées à des cas réels, ces derniers soient « propres ». Toutefois, on peut douter de la pertinence d'une méthode évaluée sur un texte incompréhensible à l'homme.

Avant de pouvoir correctement exploiter ces ressources il nous a donc fallu effectuer un pré-traitement.

Nous avons donc manuellement extrait 22 textes de cet ensemble. Ces textes ont été choisis dans les portions les moins bruitées du corpus original. Ensuite chaque texte a été relu et nettoyé du bruit considéré comme « non naturel ». Nous avons donc corrigé les problèmes pouvant provenir du procédé de récupération et d'édition du corpus original (phrases vides ou seulement composées de symboles de ponctuation, phrase répétées, etc.), mais nous avons conservé les fautes d'orthographe et de syntaxe qui elles sont probablement issues de la production du document par l'auteur et donc peuvent être considérées comme « naturelles ».

Le corpus que nous obtenons au final se compose donc de 22 textes pour un total de 1 895 phrases et 54 551 mots (table 5.2). Ce qui est très peu si l'on considère la taille du corpus

original, mais il faut prendre en considération que ce travail d'extraction et de nettoyage s'est fait manuellement. Lire, sélectionner puis extraire et nettoyer ces 22 textes est une tâche longue et pénible pour une personne compte tenu de la nature du corpus.
Si ce corpus correspond donc plus aux contraintes de nos expériences que l'original, il a toutefois un gros défaut : son manque de variété. En effet, ces 22 textes sont tous des discours politiques. Ils traitent de sujet différents certes, mais sont représentatifs d'un style particulier d'écriture. Cette uniformité dans le style pourrait être préjudiciable lors de l'analyse des résultats, laissant croire que les méthodes présentées sont conçues autour des textes.
C'est pourquoi nous avons ajouté à ces 22 discours politiques 28 textes journalistiques.

5.3.2 Un corpus de textes journalistiques

Afin de varier nos textes, il nous en fallait d'un genre différent du le discours politique et avec leurs frontières thématiques identifiées. Comme nous l'avons dit plus haut ce type de ressources est rare et dans un premier temps nous n'avons pas pu nous en procurer. Par chance les professeur Bestgen et Piérard, qui ont mené une expérience sur la segmentation thématique en utilisant un corpus segmenté par un collège de lycéens ([Bestgen & Piérard, 2006]), ont bien voulu nous laisser utiliser leur corpus pour notre évaluation.
[Bestgen & Piérard, 2006] ont constitué leur matériel de référence en se basant sur des articles extraits du journal Le Monde et prélevés sur la période des trois derniers mois de l'année 1995. Les textes sont de tailles variées (32 textes répartis en 4 catégories de 200-499 mots, 500-999 mots, 1000-1499 mots et 1500-2500 mots, soit 8 par catégorie) et traitent de sujets divers.
Chaque texte a été segmenté par 15 juges, des lycéens, et seules les phrases considérées comme des frontières thématiques par 8 juges ou plus sont identifiées comme des frontières de références. Nous sommes donc là devant une référence consensuelle. Les lycéens constituent une population d'individus ayant une maîtrise suffisante de la langue pour que leurs choix soient considérés comme crédibles. Le fait qu'il faille que plus de la moitié des juges soient d'accord pour chaque frontière renforce la crédibilité des références produites. Les lycéens ne sont certes ni des experts en linguistique ni les auteurs des textes qu'ils segmentent, mais leur avis représente la vision qu'un profane pourrait avoir et non la vision d'un spécialiste. Cette « vision de profane » est parfois plus représentative de la manière dont la langue est réellement utilisée et perçue que celle qui nous est donnée par les linguistes.
Nous avons donc utilisé 28 des 32 textes du corpus de [Bestgen & Piérard, 2006], 4 textes

ayant été écartés uniquement pour des raisons techniques, pour un total de 963 phrases et 24 930 mots (table 5.3).

Notre corpus d'expérimentation se compose donc d'un total de 50 textes pour 2 858 phrases et 79 481 mots. Les textes sont de type article journalistique et discours politique et traitent de sujet très variés (allant de la politique économique au simple fait divers). Les références qui serviront à évaluer les résultats des expériences menées sur ces textes sont des deux natures évoquées plus haut, à savoir des références d'experts et consensuelles. Nous regretterons que la répartition entre les deux origines soit si inégale, mais nous devons composer avec les ressources à notre disposition.

Identifiant	Nombre de mots	Nombre de phrases
1	617	22
2	3042	100
3	2767	92
4	1028	40
5	4532	157
6	5348	212
7	1841	47
8	1927	74
9	1789	53
10	1389	31
11	2309	81
12	7193	211
13	6097	305
14	1417	57
15	3195	79
16	1995	60
17	558	16
18	696	25
19	678	26
20	1388	57
21	3127	110
22	1618	40

TABLE 5.2 – Les textes issu du corpus de DEFT'06 en chiffres

5.4 Le choix d'un algorithme de comparaison : C99

Plutôt que de d'évaluer simplement les résultats de notre méthode et de ses variantes avec les mesures présentées plus haut, nous avons fait le choix de comparer nos résultats

Identifiant	Nombre de mots	Nombre de phrases
126	945	34
1185	283	13
1382	422	16
1768	1409	47
1771	1413	57
2005	1097	51
2063	520	19
2189	588	25
2534	362	12
2986	484	19
3591	1251	38
3805	748	38
3819	423	15
3920	437	14
3927	615	29
4240	580	17
4342	368	15
4465	579	18
4765	1909	113
5258	1521	58
5603	921	30
5995	971	31
6123	1142	43
6170	382	12
6198	1451	54
6211	1496	52
647	1101	43
731	1512	50

TABLE 5.3 – Les textes issu du corpus de Bestgen en chiffres

avec ceux obtenus par un autre algorithme : c99 ([Choi, 2000]). Cette section, en plus de présenter de manière plus détaillée que l'état de l'art l'algorithme c99, s'attache à présenter les raisons de ce choix ainsi que les conditions spécifiques de l'expérience que nous avons menée avec cet algorithme et évidemment les résultats obtenus.

5.4.1 L'algorithme c99

L'algorithme c99 est un algorithme non-supervisé qui s'appuie sur des calculs de similarités entre phrases. Il prend en entrée un texte segmenté en phrase, débarrassé de ses mots outils et de sa ponctuation et les mots conservés sont réduits à leur racine au travers d'un stemming ou d'une lemmatisation.
Une fois le texte pré-traité l'algorithme construit une matrice des fréquences des termes pour les phrases du texte, comme dans un modèle de Salton classique (comme décrit au chapitre 3). Avec ces données, c99 construit une matrice de similarité entre phrases. Toutes les phrases du texte sont donc comparées deux à deux par une mesure de cosinus et les similarités entre phrases stockées dans un tableau bi-dimensionnel (figure 5.1).
On notera que cette matrice est symétrique.
Ensuite, plutôt que d'utiliser la matrice de similarité telle quelle, l'algorithme construit une nouvelle matrice, appelée matrice de rang. Cette matrice est un classement local de chaque case de la matrice de similarité vis à vis de ses cases voisines au sein d'un masque défini par l'utilisateur. Chaque paire de phrases se voit donc attribuer un rang qui correspond au nombre de cases (donc de paires) ayant un score de similarité inférieur à celui de la case testée au sein du masque. Ce rang est normalisé par le nombre de cases réellement présentes dans le masque pour éviter les effets de bord (figure 5.2).

$$rang = \frac{Nombre\ d'éléments\ ayant\ une\ similarité\ inférieure\ dans\ le\ masque}{Nombre\ d'éléments\ réellement\ présents\ dans\ le\ masque} \tag{5.7}$$

Les segments thématiques sont ensuite identifiés via un processus de clustering inspiré de l'algorithme de maximisation de [Reynar, 1998].
Lors de nos expériences nous utiliserons le programme fourni par Choi lui-même sur son site `http://myweb.tiscali.co.uk/freddyyychoi/`. Ce programme JAVA est la version 1.3, version améliorée et optimisée de la méthode originale qui intègre un pré-traitement par LSA (présenté dans [Choi *et al.*, 2001]).

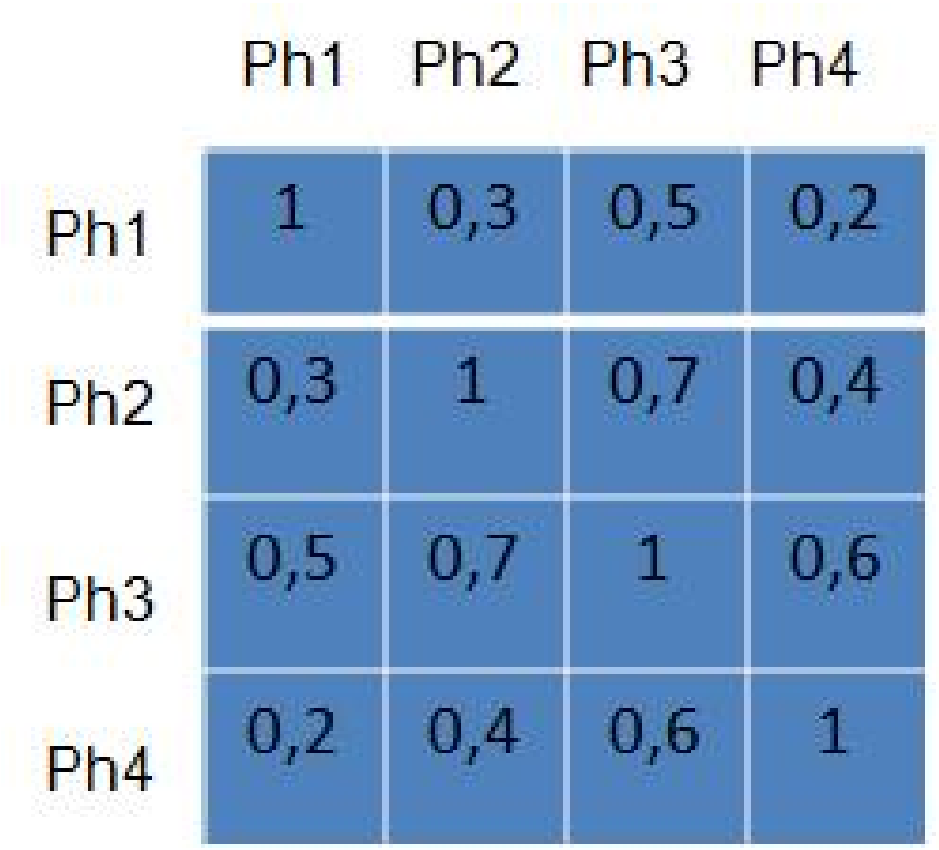

FIGURE 5.1 – Exemple de matrice de similarité

5.4.2 Pourquoi c99 ?

Plusieurs raisons nous ont poussés à choisir c99 comme comparaison de référence pour nos expériences. On peut toutefois s'arrêter sur les trois principales :

- C99 est une méthode de segmentation thématique non-supervisée comme celles que nous présentons dans cette thèse. Il était indispensable que la méthode avec laquelle nous nous mettons en concurrence soit de même type que les nôtres pour que la comparaison soit valable.
- C99 est une méthode basée sur des calculs de similarité entre phrases. Notre méthode principale s'appuie sur des calculs de distance, mais ces deux notions sont très proches (une distance pouvant être vue comme une dissimilarité). Encore une fois nous voulions comparer notre approche avec une approche ayant un maximum de points communs afin de pouvoir insister sur les différences.
- C99 est actuellement reconnu comme un des meilleurs, si ce n'est le meilleurs, algorithme de segmentation thématique par la communauté ([Bestgen & Piérard, 2006], [Sitbon & Bellot, 2004]).

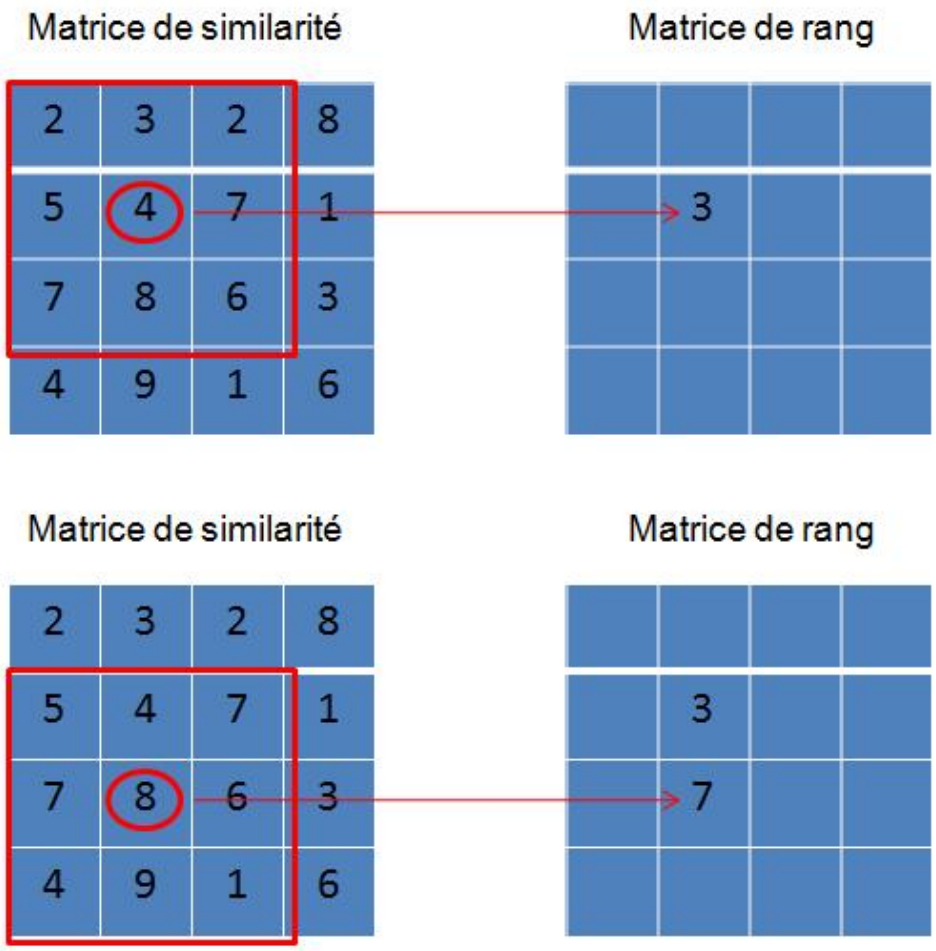

FIGURE 5.2 – Exemple simplifié de construction d'une matrice de rang

C99 était donc selon nous le meilleur candidat pour éprouver à la fois notre corpus et notre méthode.

5.4.3 Préparation du corpus pour c99

Pour leurs expériences, [Choi, 2000] ont utilisé un corpus en anglais dont on a retiré la ponctuation et les mots dit « outils » [43] et qui ont été ensuite traités par un algorithme de stemming.

Afin que les conditions soient les plus favorables possibles, il a été nécessaire de préparer le corpus pour c99. En effet, dans le cas de notre approche c'est le programme SYGFRAN qui s'occupe de la génération des vecteurs sémantiques et en quelque sorte du pré-traitement

43. Ces mots sont ceux qui participent à la compréhension du texte sur le plan syntaxique, mais pas sémantique. On peut citer parmi eux les déterminants, les pronoms, les conjonctions, etc.

du corpus. Comme nous voulions une compétition équitable, nous avons fait en sorte que le corpus soit le plus proche possible des conditions dans lesquelles [Choi *et al.*, 2001] ont mené leur expériences.
Chaque texte a donc été lemmatisé en utilisant TreeTagger ([Schmid, 1994]) et nous avons retiré des textes tous les mots outils en nous basant sur la catégorie syntaxique des mots fournie par TreeTagger. Après un tel traitement la phrase :

Un odieux attentat à la bombe a été commis contre la synagogue de la rue Copernic 'à Paris'.

devient :

odieux attentat bombe avoir être commettre synagogue rue Copernic ' Paris '

5.4.4 Résultats et commentaires

Dans les conditions que nous venons de décrire nous avons fait tourner l'algorithme c99 sur tous les textes de notre corpus et nous avons procédé à l'évaluation des résultats en utilisant les deux mesures présentées plus haut dans ce chapitre. Les résultats obtenus par la version 1.3 de l'application de [Choi *et al.*, 2001] sont présentés dans les tables 5.4, 5.5 et 5.6.

La première remarque que l'on peut faire lorsque l'on regarde ces résultats, c'est que c99 a une précision de 1 [44] dans de nombreux textes. Un tel score est étonnant, même en considérant la marge de tolérance de la mesure. Ces excellents résultats en terme de performance peuvent toutefois être expliqués très simplement :
Les conditions de l'expérience font que chaque première phrase de chaque texte est une frontière thématique. C'est très logiquement la frontière du premier segment thématique du texte. C99 (comme d'ailleurs la plupart des algorithmes de segmentation thématique, dont le notre) ramène toujours au moins la première phrase du texte comme frontière. De fait chaque fois que c99 ne ramène qu'une seule frontière, celle-ci est la première phrase et de fait est juste. La précision est donc très logiquement alors de 1, ce qui accroît très sensiblement le F_{Score}.
Dans ces conditions, il faut donc prendre également en considération le rappel et la mesure WindowDiff, le F_{Score} devenant moins significatif. Ainsi, si nous prenons par exemple le texte 1768 issu du corpus de Bestgen, on remarque que le F_{Score} de 0.2 est faible, et donc que l'algorithme n'a pas eu de bons résultats. Si l'on regarde les mesures plus en détails,

44. Nous rappelons qu'il s'agit là de la précision « souple » à deux phrases.

Id	F_{Score}	Précision	Rappel	WindowDiff
1	0.33	0.33	0.33	0.82
2	0.40	1	0.25	0.34
3	0.17	0.20	0.14	0.51
4	0.25	0.2	0.33	0.60
5	0.12	0.17	0.09	0.48
6	0.19	0.2	0.18	0.37
7	0.25	1	0.14	0.69
8	0.20	1	0.11	0.56
9	0.40	0.50	0.33	0.58
10	0.33	1	0.20	0.73
11	0.22	0.33	0.17	0.37
12	0.13	0.67	0.07	0.58
13	0.26	0.29	0.23	0.41
14	0.29	1	0.17	0.46
15	0.20	0.67	0.12	0.81
16	0.57	0.57	0.57	0.53
17	0.86	0.75	1	0.36
18	0.44	0.40	0.50	0.75
19	0.57	0.50	0.67	0.29
20	0.29	1	0.17	0.44
21	0.27	0.40	0.20	0.44
22	0.40	1	0.25	0.43
Moyenne simple juste DEFT'06	0.32	0.60	0.28	0.53
Moyenne pondérée juste DEFT'06	0.25	0.51	0.21	0.49

TABLE 5.4 – Résultats obtenus par c99 sur le corpus DEFT'06

nous remarquons un rappel extrêmement faible de 0.11 et une WindowDiff très élevée à 0.76 [45]. On peut en déduire que c99 n'a tout simplement pas fonctionné sur le texte, ne ramenant que la première phrase du texte. Nous devons donc prendre en considération cette spécificité pour évaluer correctement les performances de c99, mais aussi de nos méthodes plus tard.

Pour évaluer les résultats de manière globale (sur l'ensemble du corpus et sur chacun des sous corpus), nous avons calculé deux moyennes : Une moyenne dite « simple » qui considère chaque texte comme de manière équivalente dans son calcul et une moyenne « pondérée » qui pondère la valeur d'un texte dans le calcul de la moyenne en fonction de son nombre de phrases.

En observant les moyennes pondérées nous remarquons que les valeurs de précision, rappel et F_{Score} souples se dégradent sensiblement par rapport aux moyennes simples. Cette dégradation est visible sur l'ensemble du corpus et plus sensible sur le corpus DEFT'06 sur lequel c99 obtient de bien moins bons résultats. On en déduira que c99 est plus per-

45. Pour rappel WindowDiff est une mesure d'erreur, donc plus elle est élevée plus le score est mauvais.

Id	F_{Score}	Précision	Rappel	WindowDiff
126	0.40	1	0.25	0.48
1185	0.33	0.50	0.25	0.88
1382	0.40	1	0.25	0.82
1768	0.20	1	0.11	0.76
1771	0.60	0.75	0.50	0.40
2005	0.46	0.50	0.43	0.42
2063	0.50	1	0.33	1
2189	0.80	1	0.67	0.35
2534	1	1	1	0.43
2986	0.75	0.75	0.75	0.64
3591	0.60	1	0.43	0.61
3805	0.60	1	0.43	0.64
3819	0.75	1	0.60	1
3920	1	1	1	0.22
3927	0.44	1	0.29	0.63
4240	0.50	1	0.33	0.50
4342	0.67	1	0.50	0.60
4465	0.89	1	0.80	0.39
4765	0.33	0.33	0.33	0.33
5258	0.63	0.83	0.50	0.53
5603	0.57	0.67	0.50	0.60
5995	0.29	1	0.17	0.77
6123	0.22	1	0.13	0.79
6170	1	1	1	0.57
6198	0.60	1	0.43	0.37
6211	0.46	1	0.30	0.64
647	0.22	1	0.13	0.66
731	0.57	1	0.40	0.42
Moyenne simple juste Bestgen	0.56	0.90	0.45	0.59
Moyenne pondérée juste Bestgen	0.50	0.85	0.39	0.55

TABLE 5.5 – Résultats obtenus par c99 sur le corpus Bestgen

Id	F_{Score}	Précision	Rappel	WindowDiff
Moyenne simple globale	0.46	0.77	0.38	0.56
Moyenne pondérée globale	0.34	0.62	0.27	0.51

TABLE 5.6 – Résultats globaux

formant sur les textes de petite taille et que ses performances se dégradent avec la taille du texte à segmenter.

Dans le même temps, nous remarquons que la mesure WindowDiff s'améliore pour c99 avec la taille du texte. Comme les bonnes performances de c99 sur les petits textes peuvent être en partie expliquées par les conditions de l'expérience (comme nous l'avons expliqué plus haut), nous pouvons en conclure que la mesure WindowDiff est plus adaptée à l'éva-

luation sur des textes courts et le F_{Score} souple plus adapté pour les textes de grandes tailles.
Toutefois, quelque soit le corpus étudié, nous constatons que le point fort de c99 est la précision. C99 est donc un algorithme qui a tendance à sous segmenter, mais le peu de frontières qu'il ramène sont en général pertinentes.

5.5 Transeg : variation des paramètres et résultats

Nous avons présenté plusieurs méthodes dans les chapitres précédents. Une principale qui est le centre de nos travaux et plusieurs autres qui se trouvent être soit des variantes de notre approche centrale soit des algorithmes que nous souhaitons intégrer à notre approche.
Ces différentes approches incorporent de nombreux paramètres et nous les avons tous fait varier pour pouvoir procéder à une évaluation vraiment complète. Toutefois, le nombre de paramètres à prendre en compte est si important que les combinaisons sont trop nombreuses pour être toutes présentées ici (voir table 5.7 et 5.8). Nous avons donc sélectionné les combinaisons ayant les résultats les plus significatifs et écarté celle qui ne présentaient aucun intérêt (que ce soit parce que les paramètres n'avait aucune signification ou parce que les résultats n'étaient pas pertinents).
Pour les approches de détection passive des segments thématique nous avons testé trois

	EM	Weka densité	X-Mean
Estimation du nombre de classes	oui	oui	non
2 classes imposées	oui	oui	oui

TABLE 5.7 – Détection passive

algorithmes de clustering flous et fait varier le nombre de classes imposées.
Beaucoup plus de paramètres sont à prendre en considération pour notre méthode à détection active des frontières et ses diverses variantes. Nous avons mené l'expérience avec des fenêtres d'une taille de 10 et 20 phrases, ce qui correspond à des segments hypothétiques respectivement de taille 5 et 10. Nous avons fait varier le seuil de transition de $0,3$ à $0,75$ (par pas de $0,05$). Nous avons testé les deux principales distances présentées dans le chapitre 4. Enfin, nous avons testé les différentes formes que nous avons envisagées au chapitre 3. Au total nous avons donc 120 exécutions sur 50 textes. Ce qui justifie que nous ne présentions ici que les résultats digne d'intérêt.

	Transeg	Transeg+EM	Transeg+Hiér	Transeg+X-Mean
Taille de la fenêtre	10 et 20	10 et 20	10 et 20	10 et 20
Valeur du seuil	de 0,3 à 0,75	de 0,3 à 0,75	de 0,3 à 0,75	de 0,3 à 0,75
Type de distance	angulaire et concordance	angulaire et concordance	angulaire et concordance	angulaire et concordance
Forme	linéaire, exponentielle et sinusoïdale	linéaire, exponentielle et sinusoïdale	linéaire, exponentielle et sinusoïdale	linéaire, exponentielle et sinusoïdale

TABLE 5.8 – Détection active et fusion

5.5.1 Les résultats des méthodes par détection passive

Nous ne présenterons pas ici les résultats de la méthode par densité de probabilité proposé par la bibliothèque Weka ([Garner, 1995]), ces derniers sont très similaires aux résultats de EM, qui, comme nous allons le constater par la suite, ne sont pas bons.

5.5.1.1 EM : échec quelque soit le nombre de thèmes

L'algorithme EM fut le premier algorithme de clustering flou que nous avons testé sur notre tâche de segmentation thématique [46]. Comme il s'agissait là d'une première tentative sans véritable préparation, nous ne nous attendions pas à des résultats spectaculaires, mais nous espérions tout de même des pistes sur la meilleurs manière d'adapter le clustering flou à la tâche. Les tables 5.9, 5.10 et 5.11 montrent que ces espoirs étaient vains.

Lorsque l'on regarde ces résultats, on remarque une précision très faible et un rappel très élevé, ce qui a pour conséquence un F_{Score} médiocre. Si l'on combine cela avec une mesure WindowDiff à 1 sur presque chaque texte, on ne peut qu'en conclure que l'approche par clustering flou (du moins celle utilisant EM ou la méthode par densité de probabilité de Weka) ne fonctionne pas.

Nous avons obtenu des résultats très similaires que nous utilisions la méthode de détermination automatique du nombre de thèmes présentée dans le chapitre 3, ou que nous forcions le nombre de thèmes à seulement 2. En effet, si nous utilisons une méthode pour déterminer automatiquement le nombre de thèmes du texte et que nous imposons ainsi plus de deux classes à l'algorithme, ce dernier nous fournit en sortie une table dans laquelle seules deux classes se voient attribuer des probabilités significatives pour chaque phrase du texte, les autres classes étant simplement ignorées. En regardant de près ces probabilités, on remarque une alternance dans la répartition entre les deux classes conservées.

46. Avec l'algorithme par densité de probabilité qui a eu des résultats similaires.

Id	Précision	Rappel	F_{Score}	WindowDiff
1	0.14	1.00	0.25	1.00
2	0.10	1.00	0.18	1.00
3	0.10	1.00	0.18	0.99
4	0.08	1.00	0.15	1.00
5	0.09	1.00	0.16	1.00
6	0.06	1.00	0.12	1.00
7	0.16	1.00	0.27	1.00
8	0.13	1.00	0.22	1.00
9	0.12	1.00	0.22	1.00
10	0.17	1.00	0.29	1.00
11	0.08	1.00	0.15	1.00
12	0.16	1.00	0.27	1.00
13	0.09	1.00	0.16	0.99
14	0.50	0.17	0.25	0.56
15	0.33	0.94	0.48	0.91
16	0.13	1.00	0.22	1.00
17	0.20	1.00	0.33	1.00
18	0.17	1.00	0.29	1.00
19	0.18	1.00	0.30	0.86
20	0.13	1.00	0.24	1.00
21	0.12	1.00	0.22	0.98
22	0.10	1.00	0.19	1.00
Moyenne DEFT'06	0.15	0.96	0.23	0.97

TABLE 5.9 – Résultats obtenus par clustering flou EM sur le corpus DEFT'06

Les phrases de tous les textes analysés sont alternativement fortement (une probabilité supérieure à $0,95$) attribuées à une et l'autre des classes conservées.
Cette alternance quasi-systématique peut être expliquée par la très forte proximité entre les vecteurs sémantiques d'un même texte. L'algorithme se retrouvant dans l'incapacité de distinguer les phrases les unes des autres, mais devant produire au moins deux groupes, distribuerait alors les phrases de manière équitable entre ces deux groupes.
C'est cette alternance qui provoque ces étranges résultats, comme un rappel égal à 1 (ou très proche de 1), avec la majorité des textes. En attribuant alternativement chaque phrase à un thème puis à l'autre, toutes les phrases se retrouvent alors désignées comme des frontières thématiques. On ramène donc bien toute les frontières puisque l'on ramène tout le texte.

L'échec des algorithmes de clustering flou à fournir ne serait-ce qu'un début de réponse à notre problème nous a conduit à abandonner cette piste. Nous ne présenterons donc que très brièvement les résultats des tentatives de fusion entre ces algorithmes et notre approche linéaire, ces résultats étant tous aussi décevants, et ce pour les même raisons.

Id	Précision	Rappel	F_{Score}	WindowDiff
126	0.13	1.00	0.22	1.00
647	0.20	1.00	0.33	1.00
731	0.11	1.00	0.20	1.00
1185	0.33	1.00	0.50	1.00
1382	0.27	1.00	0.42	1.00
1768	0.20	1.00	0.33	1.00
1771	0.14	1.00	0.24	0.98
2005	0.18	1.00	0.31	1.00
2063	0.33	1.00	0.50	1.00
2189	0.13	1.00	0.22	1.00
2534	0.36	1.00	0.53	1.00
2986	0.22	1.00	0.36	1.00
3591	0.19	1.00	0.33	1.00
3805	0.19	1.00	0.32	1.00
3819	0.36	1.00	0.53	1.00
3920	1.00	0.20	0.33	1.00
3927	0.27	1.00	0.42	1.00
4240	0.19	1.00	0.32	1.00
4342	0.29	1.00	0.44	1.00
4465	0.29	1.00	0.45	1.00
4765	0.06	1.00	0.11	1.00
5258	0.19	1.00	0.32	1.00
5603	0.14	1.00	0.24	1.00
5995	0.20	1.00	0.33	1.00
6123	0.19	1.00	0.32	1.00
6170	0.27	1.00	0.43	1.00
6198	0.13	1.00	0.24	1.00
6211	0.20	1.00	0.33	1.00
Moyenne Bestgen	0.24	0.97	0.34	1.00

TABLE 5.10 – Résultats obtenus par clustering flou EM sur le corpus Bestgen

Id	Précision	Rappel	F_{Score}	WindowDiff
Moyenne	0.20	0.97	0.30	0.99

TABLE 5.11 – Résultats globaux obtenus par clustering flou EM

5.5.1.2 X-Mean : une bonne surprise

Nous n'avons envisagé d'utiliser l'algorithme X-Mean que tardivement dans nos travaux, les mauvais résultats des algorithmes de clustering flou que nous avions testés nous ayant découragé d'explorer cette piste. Nous nous sommes intéressés à l'algorithme X-Mean principalement parce que ce dernier propose de déterminer automatiquement le

nombre de classes de l'ensemble de données qu'il doit traiter.
Lors de nos expériences, nous avons constaté que X-Mean, lorsqu'il n'est pas contraint sur le nombre de classes, ne génère qu'une seul classe, quelque soit le texte. En cela, l'algorithme rejoint EM dont nous avons présenté les résultats plus haut. Toutefois, dès lors que nous avons imposé au moins deux classes (donc deux thèmes) les résultats ont été bien plus satisfaisants que les autres algorithmes de clustering testés.
Les résultats de X-Mean sur les textes politiques (5.12) de notre corpus sont meilleurs en

Id	Précision	Rappel	F_{Score}	WindowDiff
1	1.00	0.67	0.80	0.29
2	0.25	0.13	0.17	0.44
3	0.16	0.86	0.27	0.89
4	0.50	0.67	0.57	0.37
5	0.23	0.45	0.30	0.47
6	0.33	0.09	0.14	0.26
7	0.30	0.43	0.35	0.76
8	1.00	0.22	0.36	0.51
9	0.50	0.17	0.25	0.54
10	0.38	1.00	0.56	0.88
11	0.25	0.17	0.20	0.47
12	0.33	0.29	0.31	0.67
13	0.25	0.10	0.14	0.37
14	0.50	0.17	0.25	0.48
15	0.75	0.18	0.29	0.77
16	0.50	0.29	0.36	0.60
17	0.75	1.00	0.86	0.64
18	0.27	0.75	0.40	0.70
19	0.50	0.33	0.40	0.62
20	0.28	0.83	0.42	0.81
21	0.19	1.00	0.31	0.90
22	0.50	0.75	0.60	0.43
Moyenne Deft'06	0.44	0.48	0.45	0.58

TABLE 5.12 – Résultats obtenus par clustering X-Mean sur le corpus DEFT'06

terme de rappel que ceux de c99 et la précision moyenne de X-Mean est certes inférieure, mais au final le F_{Score} est supérieur à celui de c99 sur le même corpus. X-Mean est également moins performant que c99 pour ce qui est de la mesure WindowDiff, mais de peu.
Ces résultats globaux doivent toutefois être tempérés par l'irrégularité des résultats spécifiques. Si X-Mean a effectivement de bons résultats sur certains textes (comme les textes 17 et 22 par exemple), il a aussi de mauvais résultats dans d'autres. Si on regarde de près les textes où X-Mean a de bonnes performances, on constate que ce sont des textes dont les segments thématiques sont très différents les uns des autres, comme par exemple des comptes rendus de conseils des ministres durant lesquels le porte parole du gouvernement

énonce les sujets évoqués.
C'est une des spécificités des méthodes à détection passive que nous retrouvons là, ce sont des approches efficaces pour segmenter des textes abordant des sujets très variés et distincts les un des autres.
Les résultats que X-Mean obtient sur le corpus de textes journalistiques de Bestgen sont

Id	Précision	Rappel	F_{Score}	WindowDiff
126	0.27	1.00	0.42	0.90
647	0.33	0.25	0.29	0.76
731	0.13	0.40	0.19	0.87
1185	1.00	0.75	0.86	1.00
1382	0.67	0.50	0.57	0.55
1768	0.70	0.78	0.74	0.64
1771	0.24	0.67	0.35	0.69
2005	1.00	0.29	0.44	0.61
2063	0.60	1.00	0.75	0.79
2189	1.00	0.67	0.80	0.60
2534	1.00	0.50	0.67	1.00
2986	0.44	1.00	0.62	0.79
3591	0.33	0.86	0.48	0.85
3805	0.37	1.00	0.54	0.88
3819	1.00	0.40	0.57	1.00
3920	1.00	0.20	0.33	1.00
3927	1.00	0.14	0.25	0.83
4240	0.60	1.00	0.75	0.50
4342	1.00	0.50	0.67	0.60
4465	1.00	0.60	0.75	0.92
4765	1.00	0.33	0.50	0.19
5258	1.00	0.30	0.46	0.57
5603	0.33	0.25	0.29	0.60
5995	0.36	0.83	0.50	0.85
6123	1.00	0.25	0.40	0.66
6170	1.00	0.67	0.80	1.00
6198	0.24	1.00	0.39	0.92
6211	1.00	0.20	0.33	0.72
Moyenne Bestgen	0.70	0.58	0.52	0.76

TABLE 5.13 – Résultats obtenus par clustering X-Mean sur le corpus Bestgen

eux étonnants. Nous avons globalement un bon F_{Score} (et donc une précision et un rappel eux aussi satisfaisants) mais la mesure WindowDiff est elle très décevante.
Le bon F_{Score} peut être expliqué par un score de précision de 1 sur de nombreux textes [47], mais cette bonne précision est tout de même combinée à des scores de rappel élevés sur de

47. Nous avons déjà expliqué les raisons de cette bonne précision pour d'autre méthodes testées ici, ce sont les mêmes pour X-Mean.

nombreux textes. L'explication de cette différence entre un bon F_{Score} et une mesure WindowDiff décevante est que X-Mean ne propose pas ou très peu de frontières proches des frontières de références. Les frontières ramenées par X-Mean sont probablement toujours à une ou deux phrases de distance de la frontière de référence et rarement exactement dessus. De fait, le F_{Score} souple les considère comme juste, mais WindowDiff détectera de nombreuses erreurs.

Des différentes approches de détection passive par clustering que nous avons envisagée,

Id	Précision	Rappel	F_{Score}	WindowDiff
Moyenne	0.59	0.54	0.46	0.68

TABLE 5.14 – Résultats globaux obtenus par clustering X-Mean

X-Mean est celle qui a donné les meilleurs résultats. Nous regrettons de n'avoir exploré cette piste que sur la fin de nos travaux, à l'évidence l'algorithme ne peut être utilisé de manière brute comme nous l'avons fait, le paradoxe d'un F_{Score} souple satisfaisant et d'un mauvais WindowDiff en étant la preuve. Nous allons voir par la suite que la simple combinaison de X-Mean avec notre méthode linéaire par un système de vote n'est pas suffisante pour corriger ses faiblesses. Ainsi, si la piste est intéressante, elle doit encore être explorée.

5.5.2 Les résultats de l'approche par détection active et de ses variantes

Notre approche linéaire par calcul de distance entre segment hypothétique est au cœur de Transeg. Comme nous l'avons déjà dit précédemment, elle demande de nombreux paramètres. Dans les premières phases de nos travaux, ces paramètres ont été choisis arbitrairement, en nous basant sur ce qui se fait dans la littérature et sur notre intuition quand au comportement du texte.

Bien entendu, ce type de démarche n'est absolument pas satisfaisante d'un point de vue scientifique. Nous avons donc décidé de vérifier ces « intuitions » en faisant varier tous les paramètres. Nous ne présenterons ici que les configurations et les résultats les plus pertinents, la présentation de l'ensemble des possibilités étant bien trop volumineuse pour être d'un quelconque intérêt ici.

5.5.2.1 Taille de fenêtre, valeur de seuil, forme et distance pour la recherche de zone de transition

La taille de la fenêtre :

Lorsque nous avons développé notre méthode nous avons estimé qu'un segment thématique devait faire autour de 10 phrases. Cette estimation se basait sur les observations que nous avions faites sur l'ensemble des corpus fournis pour DEFT'06 (scientifique, politique et juridique) où la taille moyenne d'un segment thématique était de 10, 12. Nous avons donc choisi une fenêtre de taille 20 phrases afin qu'elle inclut en moyenne deux segments thématiques.

Lorsque nous avons travaillé sur le corpus de texte journalistique de Bestgen, nous nous sommes rendu compte que les segments thématiques étaient en moyenne beaucoup plus petits, la taille de 20 phrases étant trop importante nous avons divisé cette taille par deux et recommencé nos expériences sur l'ensemble des textes avec cette nouvelle taille [48]. Etonnamment les résultats se sont améliorés non seulement pour le corpus journalistique, mais aussi pour le corpus politique.

Les courbes présentées dans les figures 5.3 et 5.4 décrivent l'évolution du F_{Score} et de

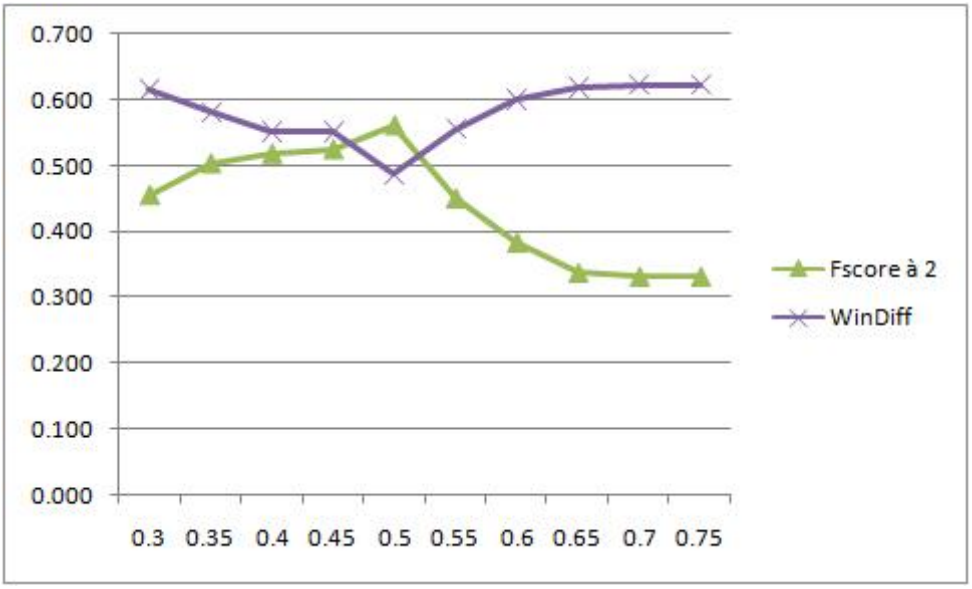

FIGURE 5.3 – Courbe pour une fonction de forme affine avec distance de concordance et fenêtre de 10

WindowDiff en fonction du seuil choisi [49]. On constate que les résultats sont sensiblement meilleurs pour une fenêtre de 10 phrases.

Il semblerait que dans la mesure où notre approche se focalise sur la nature de la frontière

48. Une taille de fenêtre de 10 phrases implique une taille théorique des segments thématiques de 5 phrases.

49. Nous discuterons du choix du seuil par la suite.

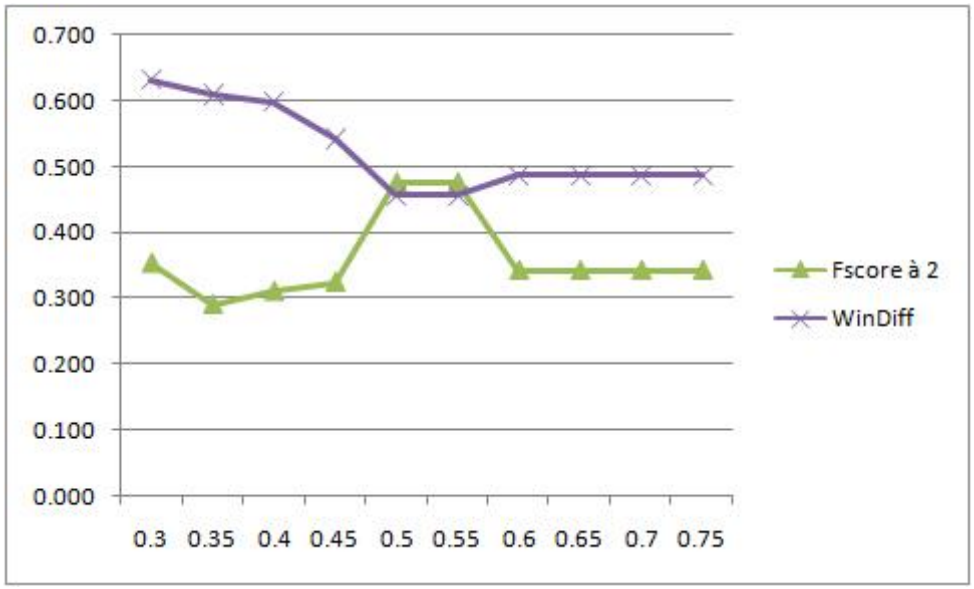

FIGURE 5.4 – Courbe pour une fonction de forme affine avec distance de concordance et fenêtre de 20

et plus spécifiquement sur les propriétés de la première phrase d'un segment thématique, elle n'ait pas besoin de beaucoup d'informations sur le segment précédent. Au contraire cette information semble être préjudiciable, créant plus de bruit qu'autre chose.

Le choix d'un seuil :
Au cours de nos travaux nous avons travaillé principalement avec des seuils autour de $0,5$[50]. Ce choix n'était pas du au hasard, mais à l'intuition que la valeur $0,5$ étant la moitié de la distance maximum pouvant séparer deux vecteurs sémantiques, les vecteurs proches les uns des autres devaient se situer en dessous de cette valeur, les vecteurs éloignés au dessus.
Nos expériences nous ont confirmé cette hypothèse.

Les courbes présentées dans les figures 5.3, 5.4, 5.5.2.1 et 5.5.2.1 montrent bien l'augmentation du F_{Score} et la baisse parallèle de WindowDiff autour de $0,5$. Et si dans plusieurs cas nous avons une zone idéale entre $0,5$ et $0,6$, le cas où nous avons les meilleurs résultats (figure 5.3) trouve son maximum (ou minimum pour WindowDiff) pour la valeur de $0,5$

50. Dans le cas de distances normalisées, donc ne pouvant prendre que des valeurs comprises entre 0 et 1.

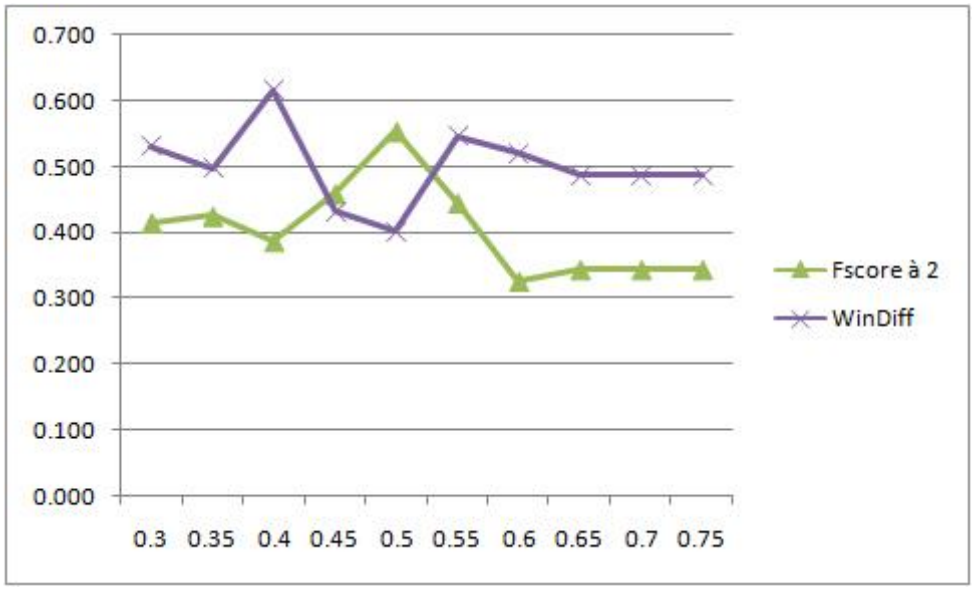

FIGURE 5.5 – Courbe pour une fonction de forme exponentielle avec distance de concordance et fenêtre de 10

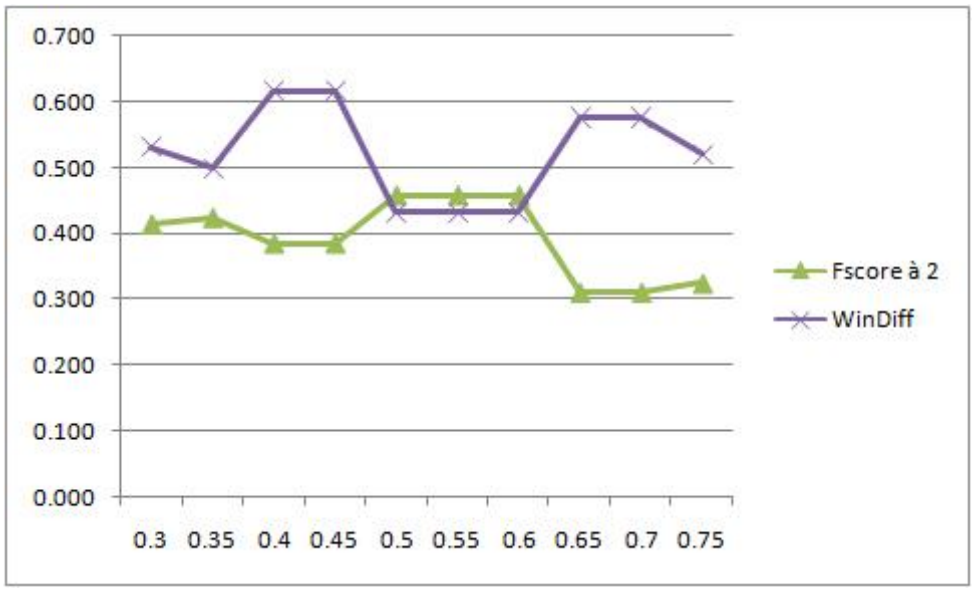

FIGURE 5.6 – Courbe pour une fonction de forme sinusoïdale avec distance de concordance et fenêtre de 10

Quelle forme choisir :
Nous avions évoqué dans le chapitre 3 le fait que la fonction de forme sinusoïdale nous semblait la plus proche de la réalité et qu'elle avait donc intuitivement nos faveurs. Nous avons également signalé que nous nous étions trompés et que les expériences le démontreraient.
Si nous regardons en détail les figures 5.3, 5.5.2.1 et 5.5.2.1, nous constatons sans doute possible que la fonction de forme affine donne de meilleurs résultats. La « forme » d'un segment thématique serait donc plus proche d'une fonction de forme affine que de toutes autres. Si ce constat peut paraître contre intuitif au premier abord, il devient logique lorsque l'on prend en considération les volumes de textes considérés. Il parait en effet difficile de « développer une information selon une courbe sinusoïdale » sur un nombre très restreint de phrases (entre 5 et 10). Peut être que si nous nous situions à un niveau de granularité plus élevé cette forme changerait vers quelque chose de plus sinusoïdale.

La distance thématique :
Nous avons évoqué dans le chapitre 3 la notion de distance thématique sans la préciser. Dans le chapitre 4 nous avons présenté plusieurs candidates possibles pour servir de distance thématique et nous avons arrêté notre choix sur deux d'entre elles :

- La distance angulaire.
- La distance de concordance.

Nous avons également fait la démonstration du caractère plus discriminant de la distance de concordance vis à vis des autres distances en nous servant d'un petit exemple simple. Les expériences à plus grande échelle ont confirmé cette supériorité de la distance de concordance sur les autres distances dans le cadre de la segmentation thématique, comme on peut le constater en regardant les figures 5.3 et 5.5.2.1.

Il en ressort que le la configuration donnant les meilleurs résultats correspond à une fenêtre de 10 phrases pour un seuil de $0,5$ avec une fonction de forme affine et en utilisant la distance de concordance. C'est cette configuration « idéale » que nous allons étudier plus en détail.

5.5.2.2 Etude détaillée de la configuration « idéale »

Nous allons donc étudier plus en détail les résultats obtenus lorsque nous sommes dans cette configuration idéale, et ce dans les mêmes conditions que notre étude des résultats de

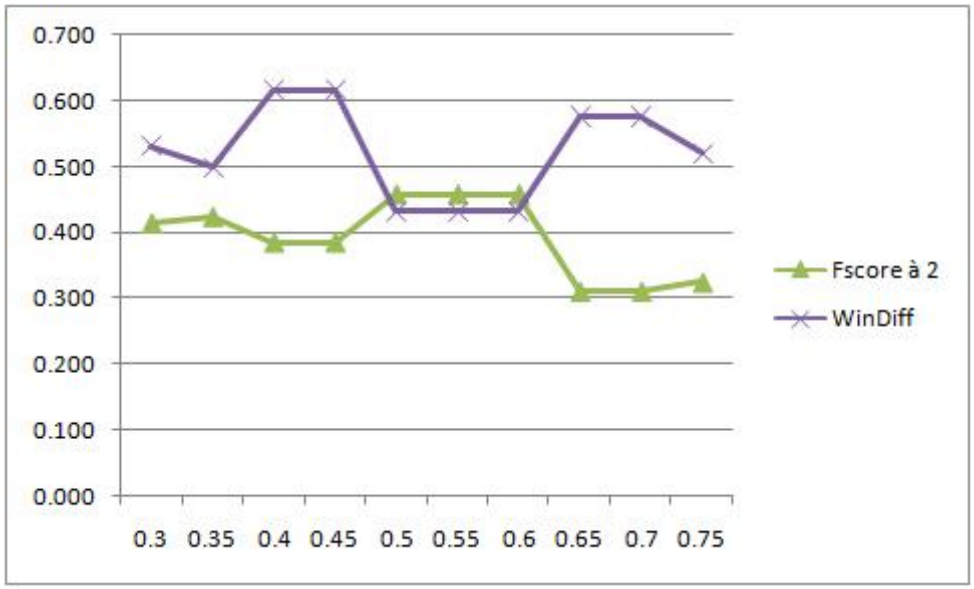

FIGURE 5.7 – Courbe pour une fonction de forme affine avec distance angulaire et fenêtre de 10

c99. Nous effectuerons donc des comparaisons avec les résultats présentés en 5.4.4 (tables 5.4, 5.5 et 5.6).

Sur le corpus politique (table 5.15)les résultats de Transeg sont relativement uniformes avec un rappel et une précision proche, même s'ils ne sont pas aussi élevé qu'on le voudrait. La mesure WindowDiff est aussi relativement basse, ce qui nous permet de déduire que nos résultats « moyens » ne sont pas dus au hasard (les deux mesures indiquant la même chose).

Comparativement à c99, Transeg est plus performant en terme de rappel et sur la mesure WindowDiff, mais moins sur la précision. Cela s'explique par la tendance de Transeg à sur segmenter alors que c99 sous segmente. L'autre constat par rapport à c99 que l'on peut faire et que si les performances de Transeg baissent avec l'augmentation de la taille des textes, cette baisse est moins sensible que pour c99. Transeg supporte donc mieux les textes de grande taille que c99.

Les résultats de Transeg sur le corpus journalistique (table 5.16) sont eux très encourageants. Si la mesure WindowDiff est quelques peu décevante par rapport aux scores obtenus sur les discours politiques, le F_{Score} et notamment la précision sont élevés. On retrouve malgré tout la supériorité de c99 en terme de précision, alors que Transeg conserve l'avantage au niveau du rappel.

Les textes journalistiques étant plus court que les textes politiques il est normal que les résultats soient globalement meilleurs sur ce corpus que sur les discours politiques. Toutefois, nous remarquons que les résultats de c99 croissent proportionnellement plus d'un corpus à l'autre. De plus les scores de Transeg ne se dégradent pas avec l'accroissement

Id	F_{Score}	Précision	Rappel	WindowDiff
1	0.86	0.75	1.00	0.29
2	0.14	0.17	0.13	0.45
3	0.35	0.30	0.43	0.56
4	0.29	0.25	0.33	0.57
5	0.28	0.22	0.36	0.53
6	0.34	0.28	0.45	0.39
7	0.71	0.60	0.86	0.40
8	0.42	0.40	0.44	0.55
9	0.20	0.25	0.17	0.52
10	0.36	0.33	0.40	0.50
11	0.50	0.50	0.50	0.41
12	0.39	0.50	0.32	0.53
13	0.33	0.27	0.43	0.49
14	0.46	0.43	0.50	0.52
15	0.74	1.00	0.59	0.55
16	0.22	0.50	0.14	0.53
17	0.40	0.50	0.33	0.91
18	0.57	0.67	0.50	0.50
19	0.40	0.50	0.33	0.38
20	0.50	0.50	0.50	0.44
21	0.50	0.50	0.50	0.44
22	0.50	0.50	0.50	0.31
Moyenne simple juste DEFT'06	0.44	0.45	0.43	0.49
Moyenne pondérée juste DEFT'06	0.40	0.39	0.42	0.48

TABLE 5.15 – Résultats obtenus par Transeg sur le corpus DEFT'06

de la taille du texte sur ce corpus, voir s'améliorent (même si c'est de manière très marginale). Ces deux constats nous permettent de confirmer que Transeg est moins dépendant de la taille du texte que c99.

De manière plus générale (table 5.17) On remarque que Transeg est plus performant en terme de rappel que c99, mais moins efficace pour ce qui est de la précision, les deux algorithmes étant équivalents sur la mesure WindowDiff. Nous nous retrouvons donc face à des comportements typiques en fonction de l'approche :

- Les approches qui détectent passivement les frontières (comme c99) ont une tendance à la sous segmentation. C'est un comportement logique puisque ces approches tendent à regrouper les phrases, cherchant des similarités. Elle auront donc plus de chance de rater des frontières. Ces approches auront logiquement un rappel faible et une bonne précision et verront leur résultats se détériorer rapidement sur des textes de plus en plus grands.
- Les approches qui détectent activement les frontières (comme Transeg) ont une

Id	F_{Score}	Précision	Rappel	WindowDiff
126	0.57	0.67	0.50	0.38
647	0.62	0.67	0.57	0.49
731	0.27	0.40	0.20	0.68
1185	0.43	0.50	0.38	0.63
1382	0.25	0.33	0.20	0.51
1768	0.33	0.50	0.25	1.00
1771	0.86	1.00	0.75	0.45
2005	0.43	0.60	0.33	0.55
2063	0.46	0.43	0.50	0.46
2189	0.55	0.75	0.43	0.58
2534	0.67	1.00	0.50	0.64
2986	0.80	1.00	0.67	0.35
3591	0.67	1.00	0.50	1.00
3805	1.00	1.00	1.00	0.36
3819	0.50	0.60	0.43	0.73
3920	0.92	1.00	0.86	0.27
3927	0.75	1.00	0.60	0.80
4240	0.33	1.00	0.20	1.00
4342	0.73	1.00	0.57	0.50
4465	0.86	0.75	1.00	0.50
4765	0.86	1.00	0.75	0.40
5258	0.25	0.33	0.20	0.69
5603	0.21	0.15	0.33	0.50
5995	0.67	1.00	0.50	0.47
6123	0.29	0.33	0.25	0.60
6170	0.60	0.75	0.50	0.50
6198	0.43	0.50	0.38	0.63
6211	1.00	1.00	1.00	0.29
Moyenne simple juste Bestgen	0.58	0.72	0.51	0.57
Moyenne pondérée juste Bestgen	0.59	0.72	0.52	0.56

TABLE 5.16 – Résultats obtenus par Transeg sur le corpus Bestgen

Id	F_{Score}	Précision	Rappel	WindowDiff
Moyenne simple globale	0.52	0.60	0.48	0.54
Moyenne pondérée globale	0.46	0.51	0.45	0.51

TABLE 5.17 – Résultats globaux obtenus par Transeg

tendance à la sur segmentation. Dans la mesure où elles cherchent à identifier les propriétés des frontières, elles en trouveront plus que nécessaire. Ces approches auront donc un bon rappel, mais une moins bonne précision et seront plus robustes face à l'accroissement de la taille des textes.

Ce qui ressort de notre expérience, c'est que Transeg, en tant qu'approche à détection

active des frontières, favorise bien évidemment le rappel. Mais lorsque l'on utilise les bons paramètres cela ce fait sans trop détériorer la précision, produisant ainsi des résultats encourageants.

5.5.2.3 La fusion avec EM

Dans le chapitre 3, nous proposons une méthode pour hybrider les approches par clustering flou avec Transeg. Toutefois, cette proposition n'est valable que si les résultats du clustering flou sont exploitables. Or nous l'avons démontré plus haut, le clustering flou (ou du moins les algorithmes que nous avons testés) ne donne pas de résultats satisfaisant lorsqu'il est adapté à une tâche de segmentation thématique.
Nous avons tout de même tenté l'expérience et observé les résultats (figure 5.5.2.3).
Le résultat était prévisible, EM ne produisant rien d'exploitable, son hybridation avec

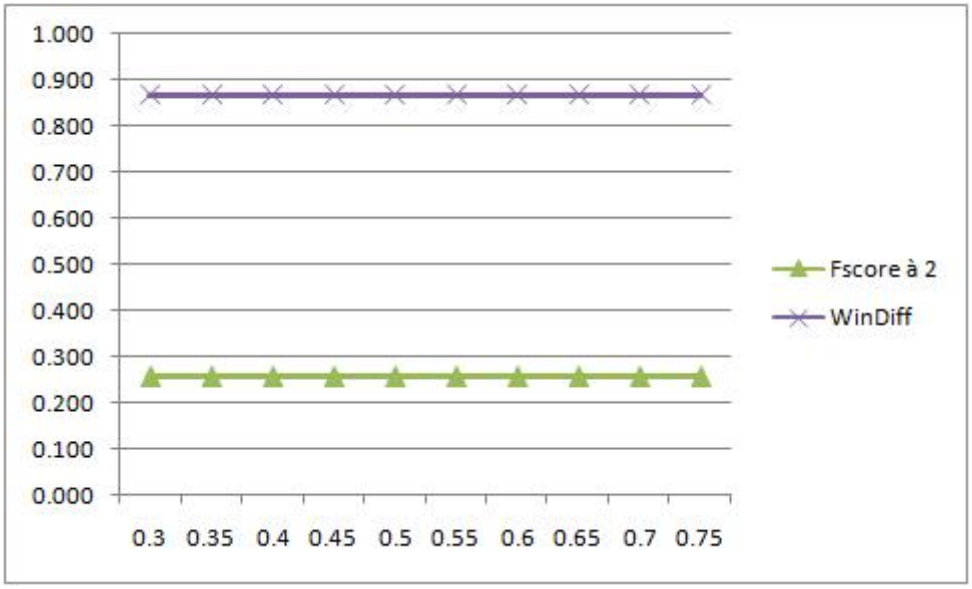

FIGURE 5.8 – Courbe pour une fonction de forme linéaire avec distance de concordance et fenêtre de 10 hybridée avec EM

Transeg produit un résultat similaire.

5.5.2.4 La fusion avec X-Mean

X-Mean fut la bonne surprise des algorithmes de clustering que nous avons testés. Bien qu'irrégulier, cet algorithme a produit des résultats intéressants et ce sans avoir été vraiment adapté à la tâche. Nous voulions voir si en l'hybridant (de manière très simple, avec un système de vote) avec Transeg, nous pouvions palier certains travers de Transeg.

Ainsi, nous aurions eu les meilleurs aspects de la détection active et de la détection passive.
Malheureusement, comme c'est souvent le cas dans ce type d'hybridation, cette combinaison a fait ressortir les mauvais aspects des deux algorithmes tout en faisant disparaître les meilleurs (figure 5.5.2.4).
Même si les résultats sont mauvais, on peut remarquer que l'hybridation permet d'atté-

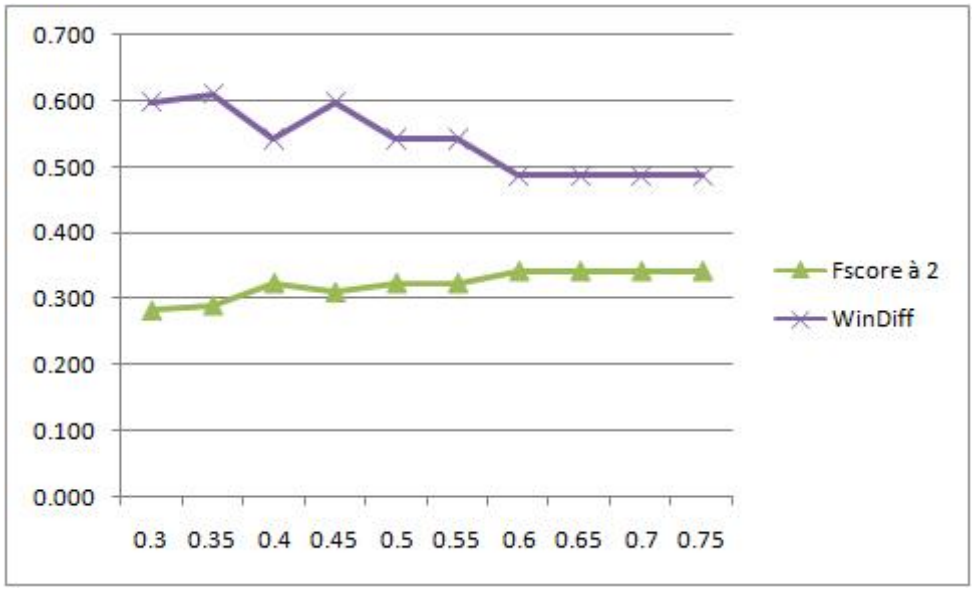

FIGURE 5.9 – Courbe pour une fonction de forme linéaire avec distance de concordance et fenêtre de 10 hybridée avec X-Mean

nuer l'effet que le choix du seuil a sur ces derniers.

5.6 Évaluation humaine et résultats

Au cour de ce chapitre, nous nous sommes intéressés à une forme « automatique » d'évaluation de nos résultats. Bien que nos références aient été produites par l'homme, nous avons utilisé des mesures statistiques pour évaluer nos résultats. Ces mesures ont l'avantage d'être objectives et reconnues par la communauté scientifique. Pourtant, tout au long de cette thèse, nous avons insisté sur le fait que la segmentation thématique est une tâche très subjective. Aussi, pour pouvoir véritablement valider notre évaluation, il faudrait que celle-ci inclut une part de « subjectivité ». Après tout, ce que nous produisons est directement ou indirectement destiné à un être humain.
De plus, comme pour l'évaluation automatique, nous voulons que notre évaluation soit comparative. Ainsi, nous avons non seulement fait évaluer par l'humain non seulement les résultats de notre approche de base, mais aussi l'algorithme c99 et les références humaines.

Nous voulons par cette démarche arriver à un résultat qui peut paraître paradoxal : évaluer le plus objectivement possible selon des critère subjectifs.

5.6.1 Protocole d'évaluation

5.6.1.1 Critères pour une évaluation humaine

La première étape dans l'élaboration de notre protocole fut de définir ce que nous voulions évaluer. Il nous fallait des critères les plus clairs et précis possibles. Nous avons donc défini deux critères pour juger de la qualité d'un segment thématique. Ces deux critères sont directement issus de notre définition du segment thématique :

- La pertinence de ses frontières.
- La cohérence interne de ses phrases.

Chacun de ces critères recevra une note attribuée par un juge humain comprise entre 0 et 5 pour chaque segment thématique du texte évalué. On aura donc l'échelle de notes suivante :

- 0 : très mauvais.
- 1 : mauvais.
- 2 : médiocre.
- 3 : acceptable.
- 4 : bon.
- 5 : très bon.

La note finale du texte dans chacun de ces critères sera leur moyenne sur l'ensemble.
La pertinence des frontières sert à évaluer si une frontière est correctement placée. Le juge donnera une note en fonction de ce qu'il pense du placement de la frontière. Est-elle à sa place ? Légèrement décalée ? Coupe-t'elle en deux un segment thématique [51] ?
La cohérence interne du segment quant à elle va permettre d'évaluer si ce dernier ne traite bien que d'un thème, où s'il en aborde plusieurs.
En plus de correspondre à la définition que nous avons faite d'un segment thématique, on peut rapprocher ces deux critères de notions bien connues lorsque l'on évalue des tâches de TALN (et même dans d'autres domaines) : le silence et le bruit. Ainsi, la pertinence peut être vue comme une mesure d'évaluation du bruit et la cohérence du silence. A la différence que ces mesures ne sont pas statistiques, mais sont des notes données par un juge humain. Nous nous rapprochons ainsi de critères statistiques en intégrant cette sub-

51. Auquel cas elle est extrêmement mal placée.

jectivité indispensable à l'évaluation de la qualité de nos résultats.

5.6.1.2 Conditions de l'évaluation

Une fois les critères définis, il nous fallait choisir les autres conditions de notre évaluation. Et parmi elles, les textes que nous allions choisir. En effet, avec 50 textes à évaluer trois fois chacun (une fois par méthode), cela ferait 150 textes à noter pour chaque juge, une tâche trop longue et trop fastidieuse pour être envisagée. Nous avons donc divisé nos texte en trois catégories :

- Une catégorie « textes courts » pour les textes de moins de 40 phrases.
- Une catégorie « textes moyens » pour les textes entre 40 et 100 phrases.
- Une catégorie « textes longs » pour les textes de plus de 100 phrases.

Puis nous avons choisi de manière aléatoire 5 textes dans chacune de ces catégories qui nous servirons de corpus d'évaluation lors de cette évaluation humaine. Ce qui fait que chaque juge aura 45 textes à évaluer ce qui est beaucoup plus acceptable.
Chaque juge se verra donc présenter chaque texte trois fois, mais ne devra pas savoir lequel est issu d'une segmentation humaine, ni quelle méthode automatique a produit le résultat. Nous souhaitons par cette approche de l'évaluation nous rapprocher le plus possible du test proposé par Alan Turing en 1950. Si un évaluateur humain ne peut pas faire la différence entre la segmentation humaine et la segmentation automatique alors nous aurons atteint notre objectif[52].

5.6.2 Évaluation sur Internet

La dernière étape, et non la moindre, de la mise en place de notre protocole d'évaluation fut le choix des évaluateurs. Pour des questions de moyens, il nous était impossible de recruter des juges et de les payer pour cette tâche. Nous avons donc décidé de faire appel à la bonne volonté des internautes pour effectuer cette tâche.
Nous avons donc proposé sur une page internet les différents textes du corpus à des évaluateurs. Dans un premier temps ces évaluateurs ont été choisi parmi des personnes proches (famille, collègues, amis, etc.), il était prévu d'étendre ce panel de juges à des personnes inconnues dans un deuxième temps, mais comme nous le verrons plus tard cela n'a pas été possible.

52. Ce qui se traduirait par des notes équivalentes, voir supérieures pour la méthode automatique.

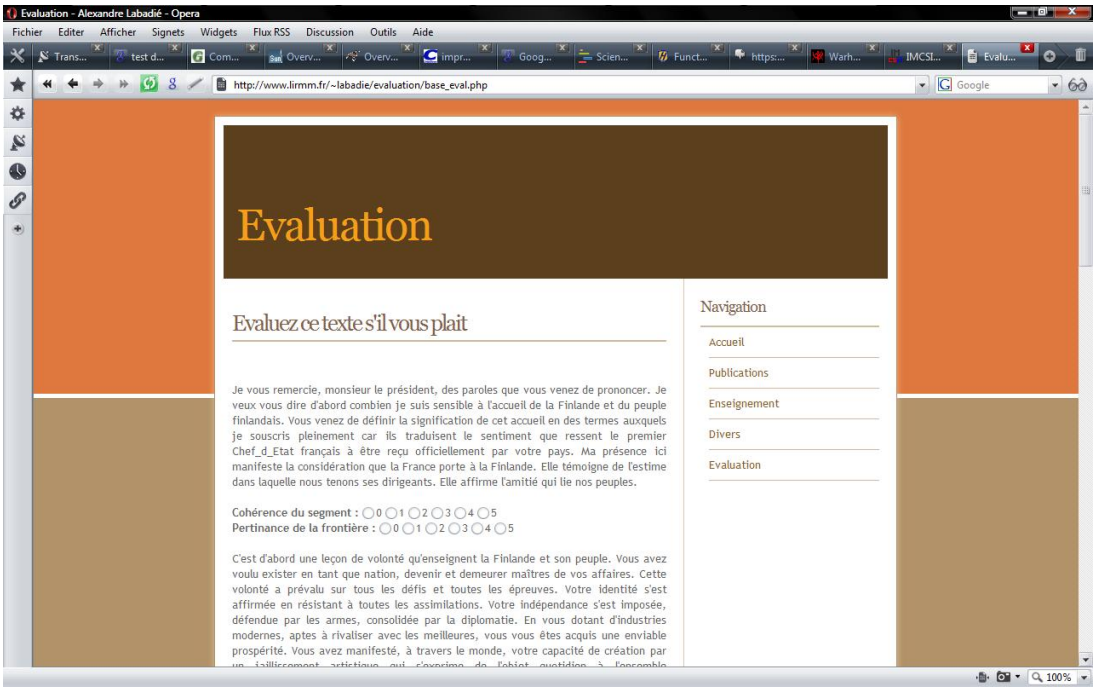

FIGURE 5.10 – Capture d'écran de la page d'évaluation

Ces juges ce sont vu proposer à chaque connexion un texte choisi aléatoirement parmi les 45 qu'il (ou elle) devait évaluer. Ainsi, un juge n'a que peu de chances de devoir évaluer le même texte segmenté par deux méthodes à la suite. L'objectif étant que le juge évalue le résultat produit pour lui-même et non pas comparativement aux autres résultats.
Sur la trentaine de personnes contactés pour la première phase de l'évaluation seulement 14 ce sont inscrites pour participer à l'évaluation. Sur ces 14 volontaires, 12 ont évalué au moins un texte et aucun n'a évalué les 45 textes sur la période de trois mois qu'a duré l'expérience.
Même simplifiée et présentée sur un support interactif, la tâche d'évaluation des résultats des différentes méthodes est particulièrement fastidieuse. Certains textes sont très longs et les sujet abordés sont rarement passionnants. Ce qui peut expliquer le manque d'engouement de notre panel de test pour notre évaluation. Cet échec nous a conduit à renoncer à élargir l'évaluation à un panel plus large. Toutefois, il nous faut relativiser cet échec car il nous a tout de même permis de récolter quelques informations et surtout il met en relief la difficulté dans le domaine du TALN de mettre en place des évaluations humaines.

5.6.3 Présentation commentée des résultats

Bien évidemment comme nous ne disposons que de résultats partiels sur un panel de juges trop restreint, nous ne pouvons tirer aucune conclusion décisive des résultats présentés ici. Par exemple seul c99 a été noté sur les 15 textes, les références humaines et Transeg n'ayant été notés que sur 14 des 15 textes.
Toutefois, nous pouvons regarder ces résultats pour en déduire des tendances et des indices, même si nous ne pouvons en tirer aucune certitude.
Les résultats présentés ici sont les moyennes des notes attribuées par les différents juges à chaque texte évalué.
Le premier constat lorsque l'on regarde les résultats obtenus par les références « humaines

Id	Pertinence des frontières	Cohérence des segments	Note globale
02	3.21	4.00	3.61
08	4.08	3.89	3.99
09	4.53	3.72	4.13
12	4.72	4.46	4.59
20	4.48	4.53	4.51
22	4.83	3.63	4.23
647	4.90	4.83	4.87
731	4.13	4.00	4.06
1185	4.56	4.67	4.61
2189	3.80	4.20	4.00
3819	4.50	4.50	4.50
4240	5.00	3.33	4.17
4765	4.20	5.00	4.60
5603	5.00	4.00	4.50
Moyenne	4.43	4.20	**4.31**

TABLE 5.18 – Résultats partiels pour les textes segmenté par l'homme

» est qu'ils sont bons (table 5.18). Même si le peu de données dont nous disposons ne nous permet pas d'émettre des hypothèses solides, nous constatons tout de même que les juges humains donnent en général de bonne notes aux résultats d'une segmentation humaine (seuls deux textes ont des notes globales inférieures à 4, et elles sont tout de même supérieures à $3,5$).
Ce constat peut paraître trivial, mais il est est indispensable pour la suite de l'évaluation. En effet, en confirmant que les références humaines que nous avons choisies sont proches des attentes d'un juge humain, nous validons les résultats de l'évaluation automatique et nous atténuons le caractère aléatoire sous entendu par la « subjectivité » de la tâche. La segmentation thématique est certes subjective et plusieurs segmentations thématiques sont probablement possibles pour un même texte, mais elle n'est pas aléatoire. Les références produites par des juges humains, que ce soient des experts ou un collège de non

spécialistes, peuvent être considérées comme fiables.

Les résultats obtenus par c99 sont assez décevants (table 5.19), et laissent supposer que

Id	Pertinence des frontières	Cohérence des segments	Note globale
02	3.42	3.80	3.61
08	2.30	2.08	2.19
09	4.10	4.08	4.09
12	4.00	3.06	3.53
20	4.00	3.67	3.83
21	3.86	3.63	3.74
22	0.00	1.67	0.83
647	0.00	1.00	0.50
731	0.00	2.60	1.30
1185	0.00	3.50	1.75
2189	0.00	2.67	1.33
3819	0.00	3.00	1.50
4240	0.00	2.00	1.00
4765	0.00	3.50	1.75
5603	0.00	1.75	0.88
Moyenne	1.30	2.73	**2.02**

TABLE 5.19 – Résultats partiels pour les textes segmenté par c99

l'algorithme ne produit pas un résultat satisfaisant pour un évaluateur humain. Mais si nous regardons en détail ces résultats, nous constatons que c99 produit en général des segments cohérents, et que ses frontières sont considérées comme pertinentes sur la majorité des textes issus du corpus politique.

Nous retrouvons là une des caractéristiques des algorithmes à détection passive : la recherche de la cohérence des segments thématiques. De plus les juges humains semblent considérer que les frontières ramenées pour les textes politiques sont cohérentes, alors qu'aucun des textes journalistiques ne s'est vu attribuer une note supérieure à 0 dans ce critère. Deux facteurs peuvent expliquer cela :

- Nous avons déjà fait remarquer que c99 a une tendance à sous segmenter. Les textes journalistiques étant plus courts que les politiques, il est possible que l'algorithme ne ramène que la première phrase de chacun d'entre eux. Or, si dans l'évaluation automatique cela lui permettait d'avoir une précision à 1 et donc de « gonfler » artificiellement ses résultats, notre évaluation humaine ne considère pas la qualité de la première frontière.
- Les meilleurs résultats sur le corpus politique peuvent s'expliquer par la nature des textes. Les discours politiques abordent en effet, en général, des thèmes plus variés et plus distincts les uns des autres (au sein du même discours) que les articles

journalistiques (qui ont tendance à ne traiter que d'un thème et de ses sous thèmes plus proches les uns des autres).

C99 est toutefois loin des score affichés par l'évaluation des résultats humains, ce qui laisse supposer qu'il ne fournit pas une solution satisfaisante pour un utilisateur humain.
Enfin, les résultats obtenus par Transeg (table 5.20) sont assez satisfaisants sur les deux

Id	Pertinence des frontières	Cohérence des segments	Note globale
02	3.00	3.00	3.00
08	3.76	3.63	3.69
09	3.78	4.08	3.93
12	2.47	3.67	3.07
20	4.00	3.83	3.92
21	3.67	2.83	3.25
22	3.00	3.75	3.38
647	4.33	4.25	4.29
731	2.25	3.38	2.81
1185	0.00	3.67	1.83
3819	4.00	4.50	4.25
4240	3.25	3.08	3.17
4765	2.90	4.17	3.54
5603	1.83	3.67	2.75
Moyenne	3.02	3.68	**3.35**

TABLE 5.20 – Résultats partiels pour les textes segmenté par Transeg

critères évalués. Encore loin derrière les résultats de la segmentation humaine, ils s'en rapprochent toutefois plus que c99. En dehors du texte 1185, Transeg a une note globale supérieure à la moyenne $2,5$ dans tous les textes du corpus. La cohérence de ses segments est en général toujours satisfaisante (supérieure à 3 sauf pour un texte), et si la pertinence de ses frontières n'est pas aussi bonne, elle reste globalement satisfaisante elle aussi.
En intégrant des principes simples et communément admis de l'organisation d'un texte, Transeg réussi à s'approcher un peu plus du résultat attendu par un utilisateur profane que c99. Cela ne signifie pas que c99 ne soit pas performant, mais peut être plus adapté à des tâches de segmentation de textes qu'à celle de segmentation **thématique** de textes.

Les résultats que nous venons de présenter ici sont partiels et insuffisants, aussi nos commentaires sur ces derniers ne peuvent être interprétés comme des affirmations ou des conclusions, mais seulement des indices sur le comportement des différentes approches testées.
Le premier et principal enseignement que nous pouvons tirer cette évaluation humaine, c'est que si ces évaluations sont indiscutablement nécessaires pour faire progresser le do-

maine, elles sont également très difficiles à mettre en place. Notre tentative de faire appel à la bonne volonté de personnes *a priori* inclinées à nous aider c'est révélée infructueuse. Il est donc logique de penser qu'étendue à un panel, certes plus vaste, d'inconnus n'ayant pas cette inclinaison, l'évaluation n'aurait pas produit plus de résultats. La conclusion de cette évaluation, certes incomplète mais encourageante, (nous avons tout de même des indices intéressants) est que pour mettre en œuvre ce type d'évaluation il faut disposer de moyens suffisants pour rémunérer (ou récompenser d'une quelconque manière) les juges.

5.7 Synthèse et conclusion

Dans ce chapitre nous avons présenté les différentes évaluations que nous avons mise en place pour tester la validité de ce que nous avons avancé dans les chapitres précédents. Dans la section 5.2, nous avons exposé les deux principales difficultés propres à l'évaluation de méthodes de segmentation thématique. Nous avons pu ainsi voir que le choix du corpus est déterminant. C'est notamment ce choix qui va déterminer si nous évaluons la qualité de la segmentation thématique des méthodes et non pas celle de la segmentation de textes. Nous avons également présenté les deux mesures utilisées lors de cette évaluation, en insistant notamment sur la nécessité de disposer de mesures souples.
Dans la section 5.3 nous avons présenté plus en détail notre corpus d'évaluation, ou, plus précisément, les deux origines de notre corpus.
Nous avons consacré la section 5.4 à l'algorithme c99. Cet algorithme, que nous avons choisi de comparer à notre approche, est considéré comme un des plus performants dans la littérature ([Bestgen & Piérard, 2006], [Sitbon & Bellot, 2004]). Toutefois, nous avons constaté que ses performances sur notre corpus ne sont pas aussi bonnes que nous pouvions l'attendre.
Dans la section 5.5, nous avons présenté l'évaluation de notre approche et de ses différentes variantes. Nous avons par exemple constaté que le clustering flou n'est pas adapté à la segmentation thématique, mais aussi que la piste du clustering n'est pas une impasse totale comme les résultats encourageants de X-Mean nous l'ont démontré. Nous y avons aussi testé les différentes configurations possibles pour notre approche et ainsi trouver les paramètres idéaux pour celle-ci. Nous avons pu ainsi vérifier que certaines de nos hypothèses, comme le seuil de transition à $0,5$, étaient valides (au moins dans le cadre de nos expériences).
Enfin, la section 5.6 nous a permis de présenter les difficultés inhérentes à une évaluation humaine et à sa mise en place. Nous y avons notamment décrit un protocole d'évaluation et sa mise en œuvre et nous avons également constaté la difficulté qu'il y a à récolter des

résultats en s'appuyant sur la seule bonne volonté des évaluateurs. Toutefois, le peu de résultats que nous avons récoltés ont été encourageants vis à vis de nos choix.

Plus généralement, il ressort de notre évaluation deux grandes questions :
- Les approches, quelqu'elles soient, que nous utilisons répondent-elles à la demande ?
Avant de savoir si une méthode de segmentation thématique (que ce soit Transeg ou une autre) répond à la demande, il faut identifier cette demande. Nous avons considéré que la segmentation thématique était à l'intention de l'humain et c'est avec cette idée en tête que nous avons développé Transeg. Alors que c99 semble plus correspondre à une demande automatique, comme la segmentation de textes [53] par exemple, qu'à ce qu'attend un utilisateur humain.
- Les mesures que nous utilisons sont elles pertinentes ?
En observant les résultats de l'évaluation automatique, nous nous sommes retrouvé face à un paradoxe, à savoir que régulièrement le F_{Score} et WindowDiff ont eu des comportements orthogonaux. Ce constat nous pousse à nous poser la question de ce qu'évaluent ces deux mesures. Le F_{Score}, en tant que moyenne lissée de la précision et du rappel, peut être vu comme l'évaluation d'un état moyen alors que WindowDiff correspondrait plutôt à une mesure des anomalies. Ce ne sont donc peut être pas les meilleures mesures pour évaluer une tâche de TALN lorsqu'elle est à l'intention de l'humain. En effet, l'humain est en général tolérant aux anomalies, mais se contente rarement d'une moyenne, il cherche plus à maximiser des critères précis, comme ceux que nous avons définis dans notre évaluation humaine.

53. Nous entendons par segmentation de textes, la tâche qui consiste à retrouver différents textes au sein d'un ensemble de textes concaténés.

6

Conclusion et perspectives

Sommaire

6.1 Synthèse

La segmentation thématique est un domaine relativement récent du TALN qui commence à émerger au début des années 1990 ([Hearst, 1993]). Comme nombre de tâches de TALN, la segmentation thématique peut être abordée de manière supervisée ou non-supervisée. Dans le chapitre 2 nous nous sommes donc attachés à présenter les différentes approches existantes de la segmentation thématique, qu'elles soient supervisées ou non-supervisées. Nous avons pu ainsi d'abord voir les limites des approches supervisées. Ces méthodes s'appuient toutes sur un même principe : la création d'un modèle à partir de textes annotés lors d'une phase d'apprentissage, puis l'utilisation de ce modèle pour retrouver les frontières thématiques d'autres textes. Les différences entre les méthodes se situent principalement au niveau de ce qu'elles apprennent dans leur modèle. Ainsi, si certaines se basent sur des distributions de mots des phrases frontières et non-frontières, d'autres se concentrent sur l'apprentissage de marqueurs linguistiques, c'est à dire d'indices spécifiques d'un changement de thème. La limite de ces méthodes réside dans le fait qu'elles ne sont performantes que sur des textes proches de ceux de leur corpus d'apprentissage, dès qu'elles sortent de ce contexte malheureusement leur performance se dégradent. Ces limites nous ont convaincus de nous orienter vers une approche non-supervisée, ce qui a justifié une étude plus approfondie de ces dernières.
Les approches non-supervisées n'utilisent que l'information qui leur est fournie par le texte et parfois des ressources généralistes comme des dictionnaires ou des thésaurus. Nous les avons classées en trois grandes catégories, pour au final nous apercevoir que toutes se

basent sur un même principe : la cohésion lexicale. Ce principe, qui veut que deux thèmes différents utilisent des champs lexicaux différents, est parfois complété avec une information syntaxique ou sémantique, mais toujours limité au niveau du mot. Notre conclusion de ce chapitre fut que tout une catégorie d'informations située au niveau de la phrase n'était pas exploitée.

Dans le chapitre 3 nous avons donc proposé un modèle exploitant l'information syntaxique, sémantique et stylistique. Nous nous sommes appuyés sur les travaux de [Chauché, 1984], qui nous ont fournis une représentation vectorielle de la phrase, celle-ci prend en compte la structure syntaxique de la phrase pour produire des vecteurs sémantiques en se basant sur un espace conceptuel défini par des linguistes ([Larousse, 1992b]). En nous basant sur cette représentation, nous avons pu construire un modèle énonçant des hypothèses stylistiques simples et communément admises pour détecter activement les frontières thématiques. Notre modèle n'est donc pas fondamentalement différent des modèles utilisant des calculs de distances (ou de similarité) que nous avons présentés dans le chapitre 2, mais il intègre deux originalités qui lui confère son caractère unique. Notre première originalité a été de ne pas nous baser uniquement sur une représentation lexicale du texte, mais de choisir une représentation de la phrase qui intègre de l'information syntaxique et sémantique. Notre seconde originalité a été d'intégrer dans nos calculs de distances une notion de structure stylistique. Comme notre modèle linéaire de segmentation thématique est, à l'évidence, une approche active de la détection des frontières, nous nous sommes intéressés à la possibilité de l'améliorer en le combinant avec une approche passive de la segmentation thématique. Plutôt que d'utiliser une approche déjà existante, nous avons exploré les possibilités qu'offrent les algorithmes de clustering dans le domaine et proposé des moyens de fusionner les résultats du clustering avec notre méthode linéaire.

Dans le chapitre 4, nous avons décrit la mise en application des éléments théoriques présentés dans le chapitre 3. Nous avons commencé par présenter la structure générale en trois phases de notre application de segmentation thématique. Notre application utilise les vecteurs sémantiques générés par l'analyseur SYGFRAN, ce qui nous a conduit à décrire assez précisément comment sont générés ces vecteurs sémantiques. En suite, nous avons exploré les différents outils à notre disposition pour comparer ces vecteurs sémantiques. Nous avons pu constater que ces outils n'étaient pas toujours adaptés à notre approche, ce qui nous a amené à développer notre propre outil : la distance de concordance. En rajoutant une notion de rang et de classement entre les différentes valeurs des vecteurs sémantiques, cette distance s'est révélée plus discriminante et donc plus satisfaisante que les autres outils à notre disposition. Nous avons enfin présenté quelques une des spécificités propres au développement de notre application de segmentation thématique.

Nous avons consacré le chapitre 5 à l'évaluation et à la validation de nos hypothèses. Nous avons d'abord évoqué les difficultés inhérentes à l'évaluation d'une tâche aussi subjective que la segmentation thématique. Nous avons ainsi pu voir que le choix du corpus de test et notamment de la manière dont les frontières thématiques de référence est primordial pour la qualité de notre évaluation. L'utilisation d'un corpus artificiel, c'est à dire composé de textes courts concaténés, est par exemple plus adapté pour évaluer de la segmentation de texte plutôt que de la segmentation **thématique** de texte. Ensuite nous nous sommes intéressés aux différentes mesures d'évaluation utilisées en segmentation thématique. Nous avons donc pu constater que le F_{Score} nécessite d'être adapté et que la mesure WindowDiff n'est pas toujours aisée à interpréter. Ensuite nous avons présenté notre corpus de test. Ce corpus se compose de textes de deux origines différentes. Une partie des textes provient du corpus politique de DEFT'O6 ([Azé *et al.*, 2006]). Ce corpus étant particulièrement bruité, il était peu exploitable. Nous avons donc sélectionné manuellement 22 textes que nous avons nettoyés afin d'en retirer une grande partie du bruit « non-naturel ». L'autre partie de notre corpus provient du corpus expérimental de [Bestgen & Piérard, 2006]. Contrairement au corpus de DEFT'06, composé de discours politiques segmentés par leur auteur (ce qui en fait un corpus d'expert), celui-ci est composé de textes du journal *Le Monde* et ses textes ont été segmentés par un collège de juges (ce qui en fait un corpus consensuel).
Avant de présenter nos résultats, nous avons présenté l'algorithme c99 de manière plus détaillé que dans notre état de l'art. Cet algorithme nous servant de référence pour comparer nos résultats à une méthode considérée comme performante dans la littérature, nous avons veillé à ce que la compétition soit la plus équitable possible. Pour ce faire, même si nous avons testé c99 sur le même corpus que notre approche, nous avons adapté ce corpus aux besoins de c99, à travers une lemmatisation et une suppression des mots considérés comme « outils ». Les résultats obtenus par c99, nous ont permis de vérifier que c99 est un algorithme favorisant la précision et dont la performance est altérée par la taille des textes. Nous avons consacré l'avant dernière partie de notre chapitre 5 à l'évaluation automatique des différentes pistes que nous avons explorées au cours de ces travaux. Nous y avons d'abord présenté les différents paramètres que nous pouvions faire varier, pour ensuite constater que les pistes que nous avons explorées dans le domaine de la détection passive ont eu des résultats décevants. A l'exception notable de l'algorithme X-Mean, qui semble offrir quelques perspectives, notre tentative d'utiliser le clustering pour faire de la segmentation thématique s'est révélé être une impasse. Ensuite, nous avons fait varier les différents paramètres de notre approche linéaire par calcul de distance. Nous avons pu constater l'impact des différents paramètre sur les résultats et ainsi déterminer la meilleurs configuration possible pour notre application. Dans cette configuration, les performances

de notre approches se sont révélées aussi bonnes, voir meilleures que celle de c99. Nous avons tout de même constaté que là où c99 se focalisait sur la précision, Transeg favorisait le rappel. De plus, les performances de Transeg se dégradent moins avec l'accroissement de la taille du texte segmenté.
Finalement, nous avons présenté notre protocole d'évaluation manuelle. Pour se faire nous avons d'abord défini deux critères, la cohérence interne des segments thématique et la pertinence de leurs frontières, pour notre évaluation. Ces deux qualités des segments thématiques ont été évaluées par des juges humains au travers d'une application internet, pour 15 textes de notre corpus et pour trois segmentation différentes : la segmentation humaine de référence, c99 et notre approche linéaire. Les résultats de cette expérience, bien qu'insuffisant pour pouvoir servir de base à des conclusions irréfutables, nous ont permis de dégager bon nombre d'indices sur le comportement de notre approche. Nous avons ainsi pu constater que Transeg produit une segmentation thématique plus proche de celle qu'attend un être humain que c99.

Les résultats satisfaisants que nous avons obtenus avec Transeg ont démontré que l'exploitation de ressources autres que la ressource lexicale est possible. Dans un domaine dominé par un usage quasi exclusif de la cohésion lexicale comme base de travail, nous avons proposé d'utiliser l'information se situant au niveau de la phrase pour segmenter le texte. En ce sens, nous nous sommes détaché du lexique, pour nous concentrer sur la phrase.

6.2 Perspective

Au cours de notre travail, nous avons pu constater qu'il reste des pistes à explorer dans la continuité du modèle que nous venons de présenter.

6.2.1 Adaptation à d'autres langues

Dans le chapitre 4, nous avons insisté sur le caractère modulaire de notre application et notamment sur sa capacité à exploiter d'autre représentation du texte que les vecteurs sémantiques générés par SYGFRAN. Notre modèle, comme notre application, nécessite une représentation vectorielle de la phrase. A l'idéal, cette représentation doit prendre en compte des informations syntaxiques et sémantiques issues d'une analyse approfondie de la phrase. Dès lors, si nous disposions d'une telle représentation pour une autre langue que le français, il est tout à fait envisageable de tester notre modèle pour cette langue.

Cela nous permettrait de vérifier que les hypothèses que nous avons émises sur la structure des segments thématiques et la nature des zones transitions, se vérifient dans d'autres langue. En effet, rien ne nous permet d'affirmer que l'organisation du discours est la même en anglais, en allemand ou en japonais qu'en français. Si notre approche se révélait être aussi efficace dans d'autres langues qu'en français, cela pourrait aider à démontrer l'existence d'une structure commune à plusieurs (voir toutes) langues dans l'organisation de la communication.

6.2.2 Construction d'une ressource multi-genres

Dans le chapitre 5, nous avons évoqué la difficulté qu'il existe à se procurer des textes annotés pour la segmentation thématique. De fait, nous n'avons pu tester nos différentes pistes que sur deux catégories de textes : des discours politiques et des articles de journaux. Ces textes, bien que différents, ont en commun leur caractère très argumentatif. Cette trop grande uniformité du style des textes sur lesquels nous avons travaillé a pu influencer notre réflexion et certains de nos choix.
Le style du texte peut avoir une grande influence sur la qualité des résultats dans d'autres domaines du TALN, comme le démontre [YM2007] dans ses travaux sur la compression automatique de phrase par exemple. Il n'y donc a pas de raisons pour que le style du texte n'influe pas sur les résultats de la segmentation thématique également. Il serait dès lors utile de construire une base de textes annotés pour la segmentation thématique qui couvre plus de styles et de genres que les discours politiques et les articles de journaux que nous avons utilisés. Nous pourrions alors étendre nos tests à des textes moins argumentatifs et plus descriptifs comme des romans par exemple. La construction d'une telle ressource pourrait ainsi servir de base à une campagne d'évaluation régulière sur la segmentation thématique, comme on en rencontre dans d'autres domaines.

6.2.3 Vers une exploitation plus systématique de la phrase

Nous l'avons constaté dans notre chapitre 2, l'information lexicale est très utilisée en segmentation thématique. Même les méthodes supervisées, à travers des modèles de distribution de mots, se basent énormément sur le niveau lexical. De fait, nous sommes dans un domaine qui extrapole des structures discursives en se basant sur le mot, sans passer par la phrase.
Faire l'économie de l'information que porte la phrase en tant que structure signifiante nous apparaît comme gâcher des ressources qui sont à notre disposition. Comme nous

l'avons évoqué dans la conclusion du chapitre 2, le passage du niveau lexical à la phrase, a été et continue à être exploré. Que ce soit des analyseurs intégrant des dépendances, des modèles de l'IA ou encore de la logique, les pistes sur l'analyse de la phrase ne manquent pas. Les vecteurs sémantiques que nous avons utilisé sont une opérationnalisation de ce passage, au moindre coût et donc au prix d'une exhaustivité de la représentation.
Le passage de la phrase au discours est donc possible sur la base de ces outils. Nous avons commencé à explorer les possibilités offertes par cette démarche pour le domaine de la segmentation thématique, mais ce passage pourrait être tenté dans d'autre comme le résumé automatique ou encore la traduction automatique. Dès lors, plutôt que de ne s'appuyer que sur des analyses génériques, on peut imaginer l'ajout d'informations spécifiques à chaque tâche à ces analyses. Dans le cadre de la segmentation thématique, ce serait par exemple des marqueurs d'articulation générique comme : « ensuite », « puis », etc.. La constitution de telles ressources pour chaque type de tâche constituerait une base intéressante de collaboration entre les « linguistes computationnels » et les linguistes « purs et durs ».

A

Glossaire

Détection active de frontière thématique : Les méthodes à détection active de frontière recherchent les frontières thématiques au travers de certaines propriétés supposées de ces frontières (rupture de chaînes lexicales, seuil de transition, etc.).

Détection passive de frontière thématique : Les méthodes à détection passive de frontière ne recherchent pas véritablement les frontières thématique, mais regroupent les phrases en segments. Les frontières apparaissent alors naturellement, « par défaut ».

Frontière thématique : Nous avons choisi de désigner comme frontière thématique la première phrase d'un segment thématique. Ce choix est notamment lié à la manière d'évaluer les méthodes de segmentation thématique.

Lemmatisation : Processus consistant a remplacer un mot par sa racine en général en se basant sur un dictionnaire.

Linguistique componentielle (ou approche sémique) : Branche de la linguistique qui postule que le sens d'un terme peut être défini par un ensemble de primitive de base. Cette idée est le prolongement des réflexions de Leibniz (1646 - 1716) qui a passé une partie de sa vie à la recherche dŠun alphabet des pensées. Si on pense quŠil peut exister un tel alphabet, il doit en exister nécessairement un qui permettrait de représenter les mots qui ne sont, après tout, que des étiquettes accolées ‘a certaines pensées. Les structuralistes, en particulier les linguistes héritiers de Leibniz comme Hjelmslev, Pottier, Greimas ou Rastier, sŠinspirent à la fois de ces idées et des théories de la phonologie pour

mettre au point lŠanalyse sémique et la théorie des primitives sémantiques qui en est une conséquence directe.

Stemming : Processus consistant à tronquer un mot pour en retrouver la racine.

TALN (Traitement Automatique du Langage Naturel ou de la Langue Naturelle) : Domaine d'étude des techniques d'analyses (compréhension) et de génération (production) automatique d'énoncés oraux ou écrits.

B

Corpus

Nous présentons ici quelques uns des textes du corpus utilisé lors de nos expériences.

Textes extrais du corpus politique de DEFT'06

Texte 1

Un odieux attentat à la bombe a été commis contre la synagogue de la rue Copernic 'à Paris'. Il a fait quatre morts, dont trois passants. Cette synagogue était, sur instruction expresse, gardée par un agent de police. J'ai exprimé par une lettre au Grand Rabbin 'Jacob KAPLAN' de France, l'indignation et la solidarité du peuple français tout entier. Au caractère criminel de l'acte s'ajoute l'écho douloureux qu'il éveille dans la communauté juive, en lui rappelant les persécutions, les déportations et les massacres systématiquement organisés par le régime hitlérien. Concernant les Français juifs qui sont des Français parmi d'autres Français, ma règle et ma préoccupation constantes sont qu'ils se sentent reconnus et traités en Français comme les autres et parmi les autres, tout en conservant, comme ils le souhaitent et comme les autres communautés françaises, leur religion et leur personnalité culturelle.

Dans cette épreuve, la communauté de tous les Français doit se resserrer, et non se diviser et se séparer. C'est pourquoi, je prescris au ministre de l'Intérieur 'Christian BONNET' d'inviter les préfets à réunir autour d'eux vendredi prochain les représentants locaux des différents cultes, des syndicats et des associations qui luttent pour la tolérance et contre le racisme, afin de témoigner entre elles de leur solidarité, et d'examiner les données locales des problèmes de sécurité. Je demande au ministre de l'Education 'Christian BEULLAC' d'inviter les recteurs à organiser, le même jour, en concertation avec les enseignants, un cours aux élèves sur le caractère pluraliste, tolérant et fraternel de la société française.

Enfin, la directive expresse a été donnée à la police sous le contrôle de la justice de poursuivre leurs investigations par tous les moyens légaux, pour découvrir les coupables, leurs complices ou leurs inspirateurs.

Il y a trois attitudes qui appellent une mise en garde de ma part : L'interprétation donnée à cet acte criminel, à l'intérieur ou à l'étranger, comme démontrant la diffusion dans le corps social français des idéaux pervers du racisme et du nazisme. De telles actions, qui sont manifestement l'oeuvre de petits groupes retranchés de la communauté nationale, n'autorisent pas une interprétation aussi basse. L'insinuation que la police ferait preuve de complaisance vis-à-vis de tels actes est injuste et condamnable. Elle s'apparente à la délation collective, de triste mémoire. Elle est d'autant plus injuste que les personnels de police et de gendarmerie ont été cruellement éprouvés ces temps derniers, dans des conditions qui appellent l'émotion et la reconnaissance. Le Gouvernement a multiplié les efforts et les moyens au cours des dernières années pour lutter contre le terrorisme. Il n'y a pas toujours été aidé. Il continue à faire confiance aux institutions démocratiques que sont la police et la justice pour assurer la sécurité et la liberté de tous les citoyens français. L'idée enfin qu'il faut répondre à la violence par la violence. Qui n'aperçoit la profondeur du piège, faisant monter la haine et appelant aux actes irréparables. La société française est une société de fraternité et de justice. C'est tous ensemble que nous ferons face aux menaces et que nous rejetterons au loin des germes hideux de l'intolérance, du terrorisme et du racisme.

Texte 4

Monsieur le Chancelier,
Mesdames et messieurs,
Vous avez entendu à l'instant le chiffre impressionnant des rencontres au sommet entre nos deux pays ! A quoi s'ajoute le chiffre encore plus impressionnant des rencontres qui m'ont permis de débattre avec le Chancelier de la République fédérale d'Allemagne. C'est dire qu'avant les grands événements qui vont se dérouler d'ici la fin de cette année, une longue période s'est déroulée, pendant laquelle nous avons appris à nous connaître, à travailler ensemble, à débattre du sort de l'Europe ainsi que des problèmes touchant à l'équilibre du monde. Nous nous sommes engagés sur la voie tracée par ceux que l'on appelle aujourd'hui « les fondateurs » de l'Europe actuelle au lendemain de la deuxième guerre mondiale. Et nous sommes allés plus loin, jusqu'à ce moment tant attendu par les Allemands qui verront leur pays enfin réunifié. On peut dire que nous avons été de bons compagnons associés dans des démarches historiques. J'allais vous dire que nous avons su surmonter bien des obstacles : ceux du passé, ceux du présent qui ont pu se résoudre

parce que nous avions les mêmes vues sur l'avenir.

L'occasion que nous offre Munich ce soir est solennelle puisque nous pouvons prononcer ces paroles à la veille de l'événement qui verra les Allemands se retrouver.Cet événement de grande ampleur aurait pu se produire dans la méfiance réciproque. Si l'on ramène tout, comme dans l'Europe traditionnelle, à des rapports de puissance et de force, c'est à l'avènement d'une Allemagne elle-même grande puissance auquel nous assistons. Nous, Français, nous sommes calés sur notre hexagone, sur la réalité solide très ancienne qui a vu sur le territoire français, ce territoire cohérent, équilibré, ramassé sur lui-même, un peuple lui-même cohérent et nous regardons l'histoire sans complexe, avec le sentiment que nous pouvons nous-mêmes l'aborder avec tous les espoirs que fournit un peuple pour lui-même. Et quel que soit le juste orgueil national que l'on puisse éprouver quand on appartient à des peuples comme le vôtre et comme le mien, nous savons bien que nous ne pouvons que dépasser ces notions si nous voulons être à la hauteur de l'histoire que nous vivons. C'est dans ce sens que je comprends la perspective et la construction de l'Europe à laquelle nous travaillons, où, il faut le dire, l'Allemagne et la France ont joué un rôle et continuent de jouer un rôle déterminant. Et, ce n'est pas au nombre d'habitants, ce n'est pas au nombre de kilomètres carré, ce n'est pas non plus au nombre des armées, ce n'est pas non plus au nombre des villes et de leur puissance, ce n'est pas non plus simplement au gré de la puissance économique que l'histoire se déterminera. Tout cela, l'histoire nous l'a donné. C'est une chance. Mais nous serions indignes de cette histoire si nous ne comprenions pas qu'après deux guerres mondiales, ce qui nous attend c'est tout autre chose. Et je ne répèterai pas ce qu'a dit le Chancelier Helmut Kohl sur les traits de cette Europe de demain matin, sur la réussite de notre entreprise, celle qui nous conduira au 1er janvier 1993, avec l'évolution des institutions, l'accroissement de la démocratie et de l'unité de vue et de décision qui réinstalleront l'Europe à la place qu'elle n'aurait jamais dû perdre, c'est-à-dire parmi les grands de ce monde. Nous avons d'autres amis, sur d'autres continents, mais il est important que l'Europe sache elle-même ce qu'elle a à faire, en respectant les amitiés, mais en se déterminant elle-même.

Nous n'allons pas en dire davantage avant que commence ce dîner dans cette admirable salle : c'est là le signe de ce que nous apporte la Bavière, dont on sait le poids, la culture, la richesse et la beauté. Au coeur de l'Europe, il y a beaucoup de leçons à apprendre, et je remercie les autorités dirigeantes, les responsables de la Bavière d'avoir voulu nous réserver ce bel accueil dans ces lieux admirables, en y apportant leur propre tempérament : chaleur, goût de vivre, et pour ce qui les concerne aussi, le goût des vastes entreprises. Demain l'Allemagne aura consacré son unité officiellement. Je souhaite que tout ce qui a été dit et fait jusqu'alors continue d'être accompli : ce n'est pas une brisure du destin,

c'est, au contraire, une confirmation. Nous formons, mesdames et messieurs, des voeux pour vous, au-delà des inévitables difficultés et des inquiétudes que pose tout avenir. Je pense que vous devez éprouver un moment d'émotion sacrée, et notre rôle à nous, c'est de la comprendre et de l'estimer. Mais, après tout cela, tout continuera. Il faudra être capable d'aborder les problèmes avec le regard aussi clair. Comme vous le voyez, à peine les événements de l'Est ont-ils provoqué cet extraordinaire succession d'événements que déjà, dans d'autres régions du monde, cela bouge de telle sorte que la paix est en jeu. Qui pourra jamais dire quand serons-nous en repos ? Jamais sans doute. Mais enfin, je l'ai dit tout à l'heure, on avance. C'est en ce sens que je forme des voeux : bonne chance à l'Allemagne, bonne chance à l'Europe. Elle se fera à force de volonté, de clairvoyance. Les conflits ne manqueront pas, ni les rivalités, ni les incompréhensions. Je ne sais même pas pourquoi je parle au futur : notre route est parsemée de ce genre de choses. Mais voyons grand. Merci à la Bavière. Bonne chance à l'Allemagne.

Textes extrais du corpus issu de l'expérimentation de Bestgen et Piérard ([Bestgen & Piérard, 2006])

Texte 126

Le choix par les épargnants d'un véhicule de placement se révèle de plus en plus compliqué. Rendu inquiet par les menaces sur le système de retraite par répartition et par le risque de chômage, l'investisseur ne peut plus se permettre de prendre des décisions sans envisager leurs conséquences sur le plan de la liquidité, de la performance et de la fiscalité des placements choisis. L'eldorado du rentier qu'auront été les années 80 appartient définitivement au passé.

La loi d'airain de l'investissement qui veut que la sécurité se paye par des performances faibles et que la recherche de gains importants se traduise par des prises de risques est redevenue réalité. L'immobilier n'est plus ce placement de toute sécurité sur lequel les Français ont construit pendant des générations leur patrimoine sans imaginer que la valeur des biens et le rendement locatif puissent baisser. La Bourse n'est pas le casino découvert au hasard des privatisations de 1986-1987 et de 1993-1995. Le krach d'octobre 1987 et les performances médiocres de la Bourse de Paris depuis plusieurs années en ont apporté la preuve. Les obligations sont, elles aussi, soumises aux turbulences des marchés, comme l'a prouvé, l'an dernier, la remontée brutale un peu partout dans le monde des taux d'intérêt. Enfin, le rendement des sicav monétaires baisse et leurs avantages fiscaux s'amenuisent.

De plus, le gouvernement s'est attaqué à certains avantages fiscaux. Les revenus des obligations et les plus-values sur actions, obligations et sicav monétaires ne bénéficient plus d'un abattement de 8 000 francs. La taxation des plus-values sur cessions d'actions ou d'obligations est alourdie. Jusqu'à présent, ces plus-values étaient exonérées d'impôt tant que le montant annuel des cessions ne dépassait pas 336 700 francs. Ce seuil sera abaissé à 200 000 francs pour 1996 (impôt payé en 1997) et 100 000 francs à compter du 1er janvier 1997 (impôt payé en 1998). Même l'assurance-vie, le placement fétiche des Français depuis plusieurs années, n'a pas été épargnée. Elle devrait pourtant rester un produit phare. Si le projet de suppression de la réduction d'impôt de 1 000 francs dont bénéficient les versements effectués en assurance-vie, ou plutôt en épargne-assurance, a provoqué un beau tapage, il ne paraît pas de nature à réduire dramatiquement les flux de capitaux dirigés sur ces placements, qui ont représenté environ 400 milliards de francs l'an dernier, en hausse de 21 % après une progression de 29 % en 1993. Les contrats restent exonérés de droits de succession ainsi que d'impôt sur les plus-values s'ils durent plus de huit ans. L'assurance-vie constitue un enjeu trop important pour les finances publiques pour que l'Etat se permette de la mettre à mal. Près de 80 % des obligations émises par l'Etat seraient absorbés par l'assurance-vie. A la direction du Trésor, chargée de lancer et de gérer les emprunts de l'Etat, on est tout à fait vigilant à ne pas tuer la poule aux oeufs d'or. Cette obsession explique la relative timidité de la réforme fiscale proposée par le gouvernement et montre à quel point est forte la dynamique d'un système d'épargne fondé sur l'inquiétude des citoyens et sur une très juteuse carotte fiscale. Les revenus capitalisés des contrats d'assurance-vie et de l'ensemble des sicav et autres fonds communs devraient pourtant, à terme, être assujettis à la contribution sociale généralisée (CSG). L'assiette et la perception d'une telle contribution est assez délicate puisqu'elle s'appliquerait à des revenus non perçus et devrait s'établir à la source, c'est-à- dire auprès des compagnies d'assurances et des banques.

A terme se profile à nouveau la création des fameux fonds de pension, un serpent de mer qui remonte à 1991. Après avoir été mis de côté par le gouvernement, le projet est revenu sur le devant de la scène à l'occasion de la présentation du projet de loi de finances pour 1996, et le premier ministre, Alain Juppé, s'est engagé sur une date proche. Les fonds de pension se veulent d'abord une réponse au problème du vieillissement de la population. Selon Denis Kessler, président de la Fédération française des sociétés d'assurances (FFSA), le nombre croissant de retraités et la baisse de la population active à partir de 2005 rendent leur création indispensable. En juin, il avait estimé qu'il serait « irresponsable » de laisser passer une année supplémentaire sans agir. Au-delà du financement des retraites, les fonds de pension auraient l'avantage de renforcer les fonds propres

des entreprises en leur assurant un financement plus stable et de drainer des capitaux frais à la Bourse de Paris, en panne d'investissement. Ses très médiocres performances, les plus mauvaises des grandes places internationales depuis 1993, s'expliquent notamment par les handicaps structurels de la place parisienne : le manque de « profondeur » et de liquidité du marché, le peu d'attrait des investisseurs français pour les actions et, ce qui en découle, la dépendance à l'égard du jugement et du sentiment des grands investisseurs institutionnels étrangers sur l'économie et les entreprises françaises.La création de fonds de pension est attendue comme une solution miracle, ce qu'elle ne serait sans doute pas dans un premier temps. La montée en puissance serait longue avant que la masse des capitaux gérés et investis à très long terme en actions ne modifient le paysage financier. Mais l'impact psychologique de la création de ces fonds de pension ne serait sans doute pas négligeable. Ce serait aussi, pour le gouvernement, un moyen de faciliter la réalisation d'un programme de privatisations dont les actionnaires sont loin d'avoir été les principaux bénéficiaires.

Texte 2005

« Nourrir la planète ». L'objectif était aussi généreux qu'ambitieux. Il fut assigné à l'Organisation pour l'alimentation et l'agriculture (FAO), il y a cinquante ans, le 16 octobre 1945, lorsque quarante-quatre pays réunis à Québec, au Canada, dans un château au style rococo, créèrent officiellement la première des institutions spécialisées des Nations unies. Un demi-siècle plus tard, force est de constater que la planète ne nourrit toujours pas la totalité de ses habitants, il s'en faut.

En Afrique subsaharienne, estiment les spécialistes, la situation nutritionnelle s'est même dégradée au cours des vingt dernières années. Et, au total, quelque 800 millions de personnes sont toujours sous-alimentées de par le monde, dont une majorité en Asie et en Afrique. Le chiffre a fléchi au fil des années ; il n'a pas baissé de façon sensible. Les progrès sont pourtant là, bien tangibles mais masqués par une explosion démographique sans précédent. Dopée par des avancées scientifiques et techniques considérables qui ont permis de tripler les rendements, la production agricole a fait mieux que coller à la croissance de la population mondiale. Elle l'a fait oublier.

D'un point de vue arithmétique, chaque individu dispose aujourd'hui de 2 700 calories/-jour contre 2 300 calories au début des années 60. Les pays en développement, dans l'ensemble, ne sont pas restés à la traîne de ce mieux incontestable. Il y a une génération, 80 % de la population du tiers-monde vivait dans des pays aux disponibilité alimentaires largement insuffisantes. Aujourd'hui, le pourcentage est inférieur à 10 %. Peut-on parier

sinon sur une accélération, du moins sur la poursuite de l'amélioration ? Autrement dit, les trente prochaines années vont-elles voir la faim et la malnutrition éradiquées ?

Un Américain, Lester Brown, directeur du Worldwatch Institute, un institut de prospective qui publie chaque année un état de la planète roboratif, s'est fait le chantre de la thèse inverse, celle d'une dégradation, en se fondant sur le ralentissement de la croissance de l'agriculture mondiale observée depuis plus d'une quinzaine d'années. Ses idées rencontrent un écho favorable dans l'opinion publique anglo-saxonne. « En 1993, fait-il observer, le produit de la pêche par habitant a baissé de 7 % par rapport à son maximum historique de 1989. A partir de 1984, l'augmentation de la production céréalière a ralenti brusquement pour retomber à un niveau inférieur au taux de croissance démographique ». Les raisons de ce tassement s'expliquent aisément, selon M. Brown : les innovations technologiques marquent le pas, la productivité stagne tandis qu'augmentent les contraintes physiques (érosion des sols, pollution atmosphérique, épuisement des nappes phréatiques, disparition des matières organiques, augmentation de la salinité des terres irriguées...). Dans ce contexte, « aucune solution n'apparaît susceptible d'inverser la tendance mondiale à la baisse de la production céréalière par habitant ». Et le prévisionniste américain de conclure : « Cela signifie que l'on ne peut plus compter sur les agriculteurs pour nourrir les nouvelles bouches que prévoient les projections démographiques. L'instauration d'un équilibre (...) dépend désormais plus des politiques de planning familial que des efforts des agriculteurs ».

La vision catastrophiste du Worldwatch Institute n'emporte pas la conviction. La baisse de la production céréalière par habitant ne résulte pas d'une quelconque « fatigue » de la terre ou d'un essoufflement du progrès technique, comme l'affirme M. Brown, mais, plus simplement, de mesures d'ajustement techniques comme le gel des terres prise par la poignée des grands pays exportateurs de céréales (Etats-Unis, Union européenne, Canada...) pour contenir l'accumulation des stocks et mettre un terme à des cours maintenus artificiellement trop bas. Deux autres causes pèsent sur la baisse de la production par habitant : le ralentissement de la croissance démographique mondiale (le nombre de bouches à nourrir croît moins rapidement qu'auparavant) et la saturation des besoins alimentaires dans les pays développés, où la consommation a atteint de tels niveaux qu'elle ne peut que plafonner. Pour les experts de la FAO, l'affaire est entendue : « Il ne paraît pas y avoir d'obstacles insurmontables en matière de ressources et de technologies au niveau mondial qui empêcheraient d'accroître les disponibilités alimentaires mondiales dans la mesure requise par la croissance de la demande réelle. (...) Une telle croissance de la production est possible même si l'on prend des mesures pour orienter l'agriculture vers un mode de production plus durable », écrivent-ils dans le rapport « Agriculture mondiale

Horizon 2010 », dont ils viennent de publier une version réactualisée.

Sur un point, en revanche, le pessimisme du Worldwatch Institute est justifié : l'épuisement progressif des ressources halieutiques. La mer est surexploitée, des espèces sont en voie de disparition, et il est vain d'espérer que les prises de poissons pourront augmenter fortement à l'avenir. Ni le recours à de nouvelles ressources, telle que l'aquaculture, ni l'évolution technologique ou des investissements accrus ne sont à même de modifier ce que les économistes appellent « les fondamentaux » : une offre mondiale de poissons inférieure à la demande.

Si la terre est capable de nourrir tous ceux qui l'habitent, pourquoi des centaines de millions d'individus continuent-ils à souffrir de malnutrition ? Pourquoi un Américain dispose-t-il de 3 600 calories quotidiennes quand un Indien doit se contenter de 2 200 calories ? Choquée de voir des agriculteurs européens détruire à intervalles réguliers des montagnes de pommes de terre ou de fruits quand on meurt de faim dans certaines régions d'Afrique, l'opinion publique est convaincue qu'il s'agit là d'un simple problème de distribution de la nourriture, d'un problème de vases communicants, en quelque sorte. C'est une vue simpliste et erronée. En réalité, si des individus ne parviennent pas à satisfaire leurs besoins alimentaires, ce ne sont pas les caprices de la nature qu'il faut incriminer. On meurt de faim parce qu'on ne dispose pas des revenus nécessaires pour assouvir ses besoins. Plutôt que de pénurie d'aliments, fait observer la FAO, mieux vaudrait parler « de pénurie de revenus ou de pouvoir d'achat, en bref, de pauvreté ou de manque de moyens donnant accès à la nourriture ». Si les plus démunis disposaient des ressources nécessaires pour cultiver leur lopin de terre ou acheter de la nourriture à autrui, le ralentissement de la croissance de la production agricole mondiale, qui divise tant les experts, n'existerait peut- être pas. C'est une perspective lointaine. Le fléau de la sous-alimentation n'est pas près de disparaître. A l'horizon 2010, pour quelque six cents millions de personnes, le souci quotidien sera toujours celui de la nourriture. Peut-on se satisfaire d'une amélioration aussi lente ?

C

La hiérarchie Larousse

CLASSE I. LE MONDE

SECTION I. LES CONCEPTS FONDAMENTAUX

1. Existence
   Existence(1), Inexistence(2), Matérialité(3), Immatérialité(4), Substance(5), Accident(6), État(7), Circonstance(8), Présence(9), Absence(10), Apparition(11), Disparition(12)
2. Identité
   Relation(13), Indépendance(14), Identité(15), Altérité(16), Ambivalence(17), Opposition(18), Substitution(19), Réciprocité(20), Ressemblance(21), Dissemblance(22), Différence(23), Uniformité(24), Diversité(25), Concordance(26), Discordance(27), Conformité(28), Non-conformité(29), Modèle(30), Imitation(31), Innovation(32), Variation(33)
3. Causalité
   Cause(34), Effet(35), Agent(36), Motif(37), But(38), Possibilité(39), Impossibilité(40), Nécessité(41), Éventualité(42), Probabilité(43), Hasard(44)

SECTION II. L'ORDRE ET LA MESURE

1. Ordre
   Ordre(45), Désordre(46), Organisation(47), Désorganisation(48), Classification(49), Méthode(50), Système(51), Règle(52), Norme(53), Normalité(54), Anormalité(55), Commencement(56), Milieu(57), Fin(58), Antériorité(59), Postériorité(60), Continuité(61), Discontinuité(62), Rang(63), Série(64), Gradation(65), Groupement(66), Inclusion(67), Exclusion(68)
2. Quantité
   Quantité(69), Mesure(70), Totalité(71), Partie(72), Unité(73), Pluralité(74), Multitude(75), Répétition(76), Complexité(77), Abondance(78), Paucité(79), Excès(80), Manque(81), Satiété(82), Égalité(83), Inégalité(84), Supériorité(85), Infériorité(86), Intensité(87), Augmentation(88), Diminution(89), Réunion(90), Séparation(91), Intégration(92), Dissociation(93), Proportion(94), Fraction(95), Reste(96), Adjonction(97), Mélange(98), Compensation(99)
3. Nombre
   Nombre(100), Zéro(101), Un(102), Deux(103), Trois(104), Quatre(105), Cinq(106), Six(107), Sept(108), Huit(109), Neuf(110), Dix(111), Douze(112), Cent(113), Mille(114), Infini(115), Calcul(116), Chiffre(117), Addition(118), Soustraction(119), Multiplication(120), Division(121), Mathématique(122)

SECTION III. L'ESPACE

1. Dimensions
   Dimension(123), Longueur(124), Largeur(125), Hauteur(126), Grosseur(127), Petitesse(128), Étroitesse(129)
2. Contours
   Extérieur(130), Intérieur(131), Bord(132), Centre(133), Contenant(134), Contenu(135), Limite(136), Revêtement(137), Barrière(138), Ouverture(139), Fermeture(140)
3. Formes
   Forme(141), Rectitude(142), Angularité(143), Courbure(144), Cercle(145), Géométrie(146)
4. Structures
   Structure(147), Ligne(148), Croix(149), Bande(150), Pointe(151), Bosse(152), Creux(153), Grain(154), Poli(155)
5. Situation
   Situation(156), Environnement(157), Intervalle(158), Soutien(159), Suspension(160), Proximité(161), Distance(162), Devant(163), Derrière(164), Dessus(165), Dessous(166), Côté(167), Droite(168), Gauche(169)

SECTION IV. LE TEMPS

1. Temps et durée
   Temps(170), Permanence(171), Durée(172), Éternité(173), Instant(174)
2. Date et chronologie
   Chronologie(175), Calendrier(176), Passé(177), Présent(178), Futur(179), Avance(180), Retard(181), Simultanéité(182), Fréquence(183), Rareté(184), Période(185), Moment(186), Saisons(187), Matinée(188), Soirée(189)
3. Évolution et histoire
   Évolution(190), Histoire(191), Événement(192), Changement(193), Nouveauté(194), Ancienneté(195), Désuétude(196)

SECTION V. LE MOUVEMENT

1. Le mouvement et ses directions
   Mouvement(197), Direction(198), Rapprochement(199), Éloignement(200), Arrivée(201), Départ(202), Entrée(203), Sortie(204), Pénétration(205), Extraction(206), Réception(207), Éjection(208), Expansion(209), Contraction(210), Montée(211), Descente(212), Saut(213), Chute(214), Rotation(215), Oscillation(216), Agitation(217), Déviation(218), Dépassement(219), Inversion(220)
2. Les forces et leurs actions
   Force(221), Traction(222), Attraction(223), Répulsion(224), Impulsion(225), Équilibre(226), Choc(227), Frottement(228), Inertie(229)

SECTION VI. LA MATIÈRE

1. Les sciences de la matière
   Chimie(230), Microphysique(231), Astronomie(232), Mécanique(233), Optique(234), Électricité(235), Magnétisme(236), Géologie(237)
2. Les propriétés de la matière
   Densité(238), Poids(239), Légèreté(240), Chaleur(241), Froid(242), Combustibilité(243), Humidité(244), Sécheresse(245), Solidité(246), Fragilité(247), Rigidité(248), Élasticité(249), Mollesse(250), Pulvérulence(251)
3. Les éléments et les matériaux

Liquide(252), Gaz(253), Bulle(254), Air(255), Feu(256), Terre(257), Minéraux(258), Minerais(259), Or(260), Argent(261), Fer(262), Bronze(263), Plomb(264), Bois(265), Verre(266), Huile(267)

4. L'environnement terrestre
   Région(268), Plaine(269), Montagne(270), Flots(271), Désert(272), Climats(273), Pluie(274), Vent(275), Nuages(276), Soleil(277), Lune(278)

SECTION VII. LA VIE

1. Le vivant
   Reproduction(279), Hérédité(280), Embryologie(281), Écologie(282), Cellule(283), Microorganismes(284)
2. Les plantes
   Botanique(285), Arbres(286), Arbustes(287), Fleurs(288), Fruits(289), Herbes et fougères(290), Champignons(291), Mousses et hépatiques(292), Algues(293), Lichens(294)
3. Les animaux
   Zoologie(295), Mammifères(296), Oiseaux(297), Poissons(298), Reptiles(299), Batraciens(300), Insectes et arachnides(301), Crustacés(302), Mollusques et petits animaux marins(303), Vers(304), Cris et bruits d animaux(305)

CLASSE II. LE MONDE

SECTION I. L'ÊTRE HUMAIN

1. Les humains
   Humains(306), Personne(307), Homme(308), Femme(309)
2. L'âge de la vie
   Vie(310), Mort(311), Âge(312), Naissance(313), Enfance(314), Jeunesse(315), Maturité(316), Vieillesse(317)

SECTION II. LE CORPS ET LA VIE

1. Le corps
   Tête(318), Membres(319), Main(320), Pied(321), Dos(322), Poitrine(323), Ventre(324), Sexe(325), Cerveau(326), Nerfs(327), Muscles(328), Os et articulations(329), Dents(330), Coeur et vaisseaux(331), Sang(332), Glandes(333), Peau(334), Pilosité(335), Tissus vivants(336)
2. Les fonctions vitales
   Nutrition(337), Digestion(338), Excrétion(339), Respiration(340), Sexualité(341), Immunité(342)

SECTION III. LE CORPS ET LES PERCEPTIONS

1. Sensation
   Sensation(343), Inconscience(344), Douleur(345)
2. La vision et le visible
   Vision(346), Troubles de la vision(347), Visibilité(348), Invisibilité(349), Lumière(350), Obscurité(351), Couleur(352), Blanc(353), Noir(354), Gris(355), Brun(356), Rouge(357), Jaune(358), Vert(359), Bleu(360), Violet(361), Polychromie(362)
3. L'audition et le son
   Audition(363), Surdité(364), Son(365), Silence(366), Bruit(367), Sifflement(368), Stridence(369), Son grave(370)
4. L'odorat et le parfum
   Odeur(371), Parfum(372)

5. Le goût
   Goût(373)
6. Le toucher
   Toucher(374)

SECTION IV. LE CORPS ET SON ÉTAT

1. La santé, l'hygiène et les maladies
   Vigueur(375), Faiblesse(376), Veille(377), Sommeil(378), Nudité(379), Propreté(380), Saleté(381), Santé(382), Maladie(383), Guérison(384), Aggravation(385), Malformation(386), Blessure(387), Tumeur(388), Empoisonnement(389), Toxicomanie(390)
2. La médecine et les soins du corps
   Médecine(391), Chirurgie(392), Soins du corps(393), Médicaments(394), Diététique(395)

SECTION V. L'ESPRIT

1. L'intelligence et la mémoire
   Intelligence(396), Sottise(397), Entendement(398), Aveuglement(399), Mémoire(400), Oubli(401), Attention(402), Inattention(403), Imagination(404), Curiosité(405), Finesse(406)
2. La connaissance et la vérité
   Savoir(407), Ignorance(408), Vérité(409), Erreur(410), Découverte(411), Recherche(412), Apprentissage(413), Enseignement(414), Éducation(415)
3. Le raisonnement
   Raisonnement(416), Affirmation(417), Négation(418), Question(419), Réponse(420), Idée(421), Principe(422), Supposition(423), Intuition(424), Comparaison(425), Contrôle(426)
4. Le jugement et les valeurs
   Jugement(427), Accord(428), Désaccord(429), Certitude(430), Incertitude(431), Surestimation(432), Sous-estimation(433), Qualité(434), Médiocrité(435), Beauté(436), Laideur(437), Importance(438), Insignifiance(439)

SECTION VI. L'AFFECTIVITÉ

1. Les caractères
   Sensibilité(440), Insensibilité(441), Optimisme(442), Pessimisme(443), Entrain(444), Paresse(445), Patience(446), Impatience(447), Calme(448), Nervosité(449), Folie(450)
2. Les dispositions d'esprit
   Enthousiasme(451), Réserve(452), Sérieux(453), Moquerie(454), Attirance(455), Aversion(456), Attente(457), Ennui(458), Surprise(459), Regret(460), Déception(461), Souci(462)
3. Les émotions
   Joie(463), Tristesse(464), Comique(465), Tragique(466), Plaisir(467), Déplaisir(468), Satisfaction(469), Insatisfaction(470), Colère(471), Peur(472), Soulagement(473), Espoir(474), Désespoir(475)

SECTION VII. LA VIE SPIRITUELLE

1. La pensée religieuse et philosiphique
   Religion(476), Théologie(477), Philosophie(478), Foi(479), Incroyance(480)
2. Le sacré et le profane
   Sacré(481), Profane(482), Sacrilège(483), Magie(484), Divination(485)
3. Les religions
   Judaïsme(486), Christianisme(487), Islam(488), Bouddhisme(489), Hindouisme(490)
4. Les cultes et les pratiques

Culte(491), Religieux et ministres des cultes(492), Lieux de culte(493), Prière(494), Prédication(495), Messe(496), Fêtes religieuses(497), Pape(498), Moines(499)

5. Les croyances
   Divinités(500), Textes sacrés(501), Dieu(502), Ange(503), Démon(504), Paradis(505), Enfer(506)

SECTION VIII. LA VOLONTÉ

1. Décision et indécision
   Volonté(507), Courage(508), Lâcheté(509), Résolution(510), Irrésolution(511), Persévérance(512), Défection(513), Obstination(514), Renonciation(515)
2. Le libre-arbitre et la nécessité
   Liberté(516), Fatalité(517), Obligation(518), Choix(519), Refus(520), Prétexte(521), Caprice(522), Désir(523), Indifférence(524), Persuasion(525), Dissuasion(526)

SECTION IX. L'ACTION

1. L'action et l'inaction
   Action(527), Réaction(528), Inaction(529), Effort(530), Repos(531)
2. Le projet et son résultat
   Intention(532), Tentative(533), Projet(534), Entreprise(535), Préparation(536), Impréparation(537), Accomplissement(538), Inaccomplissement(539), Succès(540), Échec(541)
3. Les occasions et les circonstances
   Opportunité(542), Inopportunité(543), Utilité(544), Inutilité(545), Facilité(546), Difficulté(547), Prospérité(548), Adversité(549), Sécurité(550), Danger(551), Avertissement(552), Alarme(553), Obstacle(554), Détection(555)
4. Les objectifs
   Construction(556), Destruction(557), Réparation(558), Préservation(559), Protection(560), Annulation(561)
5. La participation
   Participation(562), Aide(563), Stimulation(564), Encouragement(565), Conseil(566)
6. Les manières d'agir
   Usage(567), Habitude(568), Abus(569), Adresse(570), Maladresse(571), Prudence(572), Imprudence(573), Soin(574), Négligence(575), Rapidité(576), Lenteur(577), Ponctualité(578), Modération(579), Violence(580)

CLASSE III. LA SOCIÉTÉ

SECTION I. LE RAPPORT A L'AUTRE

1. Les comportements
   Sociabilité(581), Insociabilité(582), Compagnie(583), Solitude(584), Bonté(585), Méchanceté(586), Générosité(587), Égoïsme(588), Gratitude(589), Hospitalité(590), Inhospitalité(591), Courtoisie(592), Discourtoisie(593), Loyauté(594), Hypocrisie(595), Promesse(596), Trahison(597), Délicatesse(598), Dureté(599)
2. Les sentiments
   Amour(600), Caresse(601), Passion(602), Ressentiment(603), Amitié(604), Inimitié(605), Confiance(606), Défiance(607), Jalousie(608), Pitié(609)
3. L'image de soi
   Fierté(610), Honte(611), Modestie(612), Prétention(613), Distinction(614), Affectation(615), Simplicité(616), Ostentation(617), Timidité(618), Décence(619), Indécence(620)

SECTION II. LE RAPPORT HIÉRARCHIQUE

1. Autorité et soumission
   Autorité(621), Domination(622), Influence(623), Obéissance(624), Désobéissance(625), Respect(626), Irrespect(627), Soumission(628), Servilité(629), Résistance(630)
2. Commandement et consentement
   Commandement(631), Autorisation(632), Interdiction(633), Demande(634), Consentement(635)
3. Louange et repproche
   Louange(636), Reproche(637), Pardon(638)
4. Le prestige social
   Gloire(639), Ostracisme(640), Honneur(641), Discrédit(642), Promotion(643), Éviction(644), Ridicule(645), Noblesse(646), Roture(647), Titres(648)

SECTION III. GUERRE ET PAIX

1. Le conflit et le compromis
   Conflit(649), Guerre(650), Révolution(651), Paix(652), Compromis(653), Pacte(654)
2. Les épisodes du conflit
   Attaque(655), Défense(656), Injure(657), Coup(658), Représailles(659), Victoire(660), Défaite(661), Revanche(662)
3. La force armée
   Armée(663), Armes(664), Armement ancien(665), Manoeuvres(666), Tir(667)

SECTION IV. LA VIE COLLECTIVE

1. Société et organisation politique
   Société(668), Politique(669), Régime(670), Systèmes politiques(671), Élection(672), Représentants(673)
2. Citoyenneté
   Citoyen(674), Civisme(675), Habitant(676), Étranger(677)
3. La famille
   Famille(678), Père(679), Mère(680), Filiation(681), Mariage(682), Célibat(683), Divorce(684)
4. Les coutumes
   Coutume(685), Cérémonies(686), Fête(687), Funérailles(688), Salutations(689)

SECTION V. LA MORALE

1. La loi morale
   Morale(690), Devoir(691), Prescription(692), Honnêteté(693), Malhonnêteté(694), Mérite(695), Imperfection(696), Péché(697), Expiation(698)
2. Les vertus et les vices
   Vertu(699), Vice(700), Tempérance(701), Ascèse(702), Intempérance(703), Chasteté(704), Luxure(705), Sobriété(706), Gloutonnerie(707), Ivrognerie(708), Avarice(709), Prodigalité(710)

SECTION VI. LE DROIT

1. La justice
   Justice(711), Injustice(712), Droit(713), Tribunal(714), Plaidoirie(715), Police(716)
2. Les délits et les peines
   Vol(717), Escroquerie(718), Proxénétisme(719), Crime(720), Arrestation(721), Condamnation(722), Détention(723), Libération(724), Supplice(725)

SECTION VII. LA COMMUNICATION ET LE LANGAGE

1. Communication et dissimulation
   Communication(726), Secret(727), Tromperie(728), Mensonge(729)
2. Le signe et le sens
   Signe(730), Représentation(731), Sens(732), Non-sens(733), Intelligibilité(734), Inintelligibilité(735), Ambiguïté(736), Sous-entendu(737), Interprétation(738)
3. La langue
   Langue(739), Grammaire(740), Phrase(741), Mot(742), Nom(743), Lettre(744)
4. La parole
   Parole(745), Troubles de la parole(746), Cri(747), Interjections(748), Conversation(749), Plaisanterie(750)
5. Le discours
   Discours(751), Figures de discours(752), Rhétorique(753), Récit(754), Description(755), Résumé(756)
6. Le style
   Éloquence(757), Platitude(758), Concision(759), Prolixité(760), Grandiloquence(761)

SECTION VIII. LA COMMUNICATION ET L'INFORMATION

1. L'écrit et les médias
   Écriture(762), Imprimerie(763), Imprimé(764), Livre(765), Presse(766), Radiotélévision(767), Publicité(768)
2. Circulation et traitement de l'information
   Télécommunications(769), Correspondance(770), Enregistrement(771), Informatique(772)

SECTION IX. L'ART

1. Arts plastiques image et décor
   Peinture et dessin(773), Iconographie(774), Photographie(775), Sculpture(776), Architecture(777), Ornements(778), Art des jardins(779), Tendances artistiques(780)
2. La musique et la chanson
   Musique(781), Musiciens(782), Instruments de musique(783), Chant(784), Chanson(785)
3. Les arts du spectacle
   Danse(786), Théâtre(787), Scène(788), Poésie(789), Cinéma(790), Cirque(791)

SECTION X. LES ACTIVITÉS ÉCONOMIQUES

1. Le travail et la production
   Emploi(792), Main-d oeuvre(793), Lieu de travail(794), Salaire(795), Production(796), Improduction(797)
2. L'industrie et l'artisanat
   Énergie(798), Outils(799), Machines(800), Manutention(801), Exploitation minière(802), Pétrole(803), Pétrochimie(804), Sidérurgie(805), Travaux publics(806), Menuiserie(807), Plomberie(808), Serrurerie(809), Textile(810)
3. L'agriculture et la pêche
   Agriculture(811), Arboriculture(812), Élevage(813), Pêche(814)
4. Les transports
   Transports(815), Transports par route(816), Automobile(817), Transports par rail(818), Transports maritimes et fluviaux(819), Transports par air(820), Astronautique(821)
5. Le commerce et les biens
   Possession(822), Cession(823), Restitution(824), Paiement(825), Don(826), Commerce(827), Marchandise(828)

6. L'économie
   Richesse(829), Pauvreté(830), Prix(831), Cherté(832), Modicité(833), Gratuité(834), Dépense(835), Dette(836), Libéralisme(837), Dirigisme(838)
7. La finance
   Monnaie(839), Banque(840), Crédit(841), Bourse(842), Valeurs mobilières(843), Épargne(844), Gestion(845), Fiscalité(846)

SECTION XI. LA VIE QUOTIDIENNE

1. L'habitat
   Habitat(847), Maison(848), Urbanisme(849), Mobilier(850), Vaisselle(851), Éclairage(852), Chauffage(853), Nettoyage(854)
2. L'alimentation
   Repas(855), Gastronomie(856), Pain(857), Sucrerie(858), Boisson(859), Produits laitiers(860), Fromage(861)
3. Le vêtement et la parure
   Vêtement(862), Mode(863), Couture(864), Chaussure(865), Bijou(866), Coiffure(867)
4. Les loisirs
   Passe-temps(868), Voyage(869), Sports(870), Chasse(871), Jeux(872), Jouet(873)

Bibliographie

[Azé *et al.*, 2006] J. Azé, T. Heitz, A. Mela, A. Mezaour, P. Peinl, et Mathieu Roche. ń Présentation de DEFT'06 (DEfi Fouille de Textes) ż. *Actes de DEFT'06*, pp 3–12, 2006.

[Barzilay & Elhadad, 2000] Regina Barzilay et Michael Elhadad. 2000.

[Beeferman *et al.*, 1997] Doug Beeferman, Adam Berger, et John Lafferty. Text Segmentation Using Exponential Models. Dans les actes de Claire Cardie et Ralph Weischedel, , *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp 35–46. Association for Computational Linguistics, Somerset, New Jersey, 1997.

[Beeferman *et al.*, 1999] Doug Beeferman, Adam Berger, et John Lafferty. ń Statistical models for text segmentation ż. Dans les actes de *Machine Learning*, volume 34, pp 177–210, 1999.

[Bestgen & Piérard, 2006] Yves Bestgen et S. Piérard. ń Comment évaluer les algorithmes de segmentation automatiques ? Essai de construction d'un matériel de référence. ż. *Actes de TALN'06*, 2006.

[Brants *et al.*, 2002] T. Brants, F. Chen, et I. Tsochantaridis. ń Topic-based document segmentation with probalistic latent semantic analysis ż. *Proceedings of CIKM*, pp 211–218, 2002.

[BY1999] R. Baeza-Yates et B. Ribeiro-Neto. *Modern Information retrieval*. Addison-Wesley Longman Publishing Co., 1999.

[Celeux & Govaert, 1991] G. Celeux et G. Govaert. ń A classification EM algorithm for clustering and two stochastic versions. ż. , Inria, 1991.

[Celeux, 1985] D. Celeux, G. et Diebolt. ń The sem algorithm : a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. ż. *Computational Statistics Quarterly*, page 73Ű82, 1985.

[Charolles, 1997] M. Charolles. ń LŠ encadrement du discours - Univers, champs, domaines et espace. ż. *Cahier de recherche linguistique, 6.*, 1997.

[Chauché & Prince, 2007] Jacques Chauché et Violaine Prince. ń Classifying texts through natural language parsing and semantic filtering. ż. *Proceedings of LTC'03*, 2007.

[Chauché, 1984] Jacques Chauché. ń Un outil multidimensionnel de l'analyse du discours ż. *Proceedings of Coling'84*, pp 11–15, 1984.

[Chauché, 2001] Jacques Chauché. ń *SYGMART : Manuel de référence Version 4.0* ż, 2001.

[Chauché *et al.*, 2003] Jacques Chauché, Violaine Prince, Simon Jaillet, et M. Teisseire. ń Classification Automatique de Textes à partir de leur Analyse Syntaxico-Sémantique ż. *Actes de TALN'03*, pp 55–65, 2003.

[Choi, 2000] Fred Y. Y. Choi. ń Advances in domain independent linear text segmentation. ż. *Proceeding of NAACL-00*, pp 26–33, 2000.

[Choi *et al.*, 2001] Fred Y. Y. Choi, P. Wiemer-Hastings, et J. Moore. ń Latent Semantic Analysis for Text Segmentation ż. *Proceedings of EMNLP*, pp 109–117, 2001.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, et Harshman R.. ń Indexing by Latent Semantic Analysis. ż. *Journal of the American Society for Information Science*, pp 391–407, 1990.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, et D. B. Rubin. ń Maximum likelihood from incomplete data via the em algorithm (with discussion). ż. *Journal of the Royal Statistical Society*, page 1Ű38, 1977.

[Farzindar, 2004] Atefeh Farzindar. ń Développement d'un suystème de Résumé automatique de Textes Juridiques ż. *Actes de RECITAL'04*, 2004.

[Ferret, 2006] Olivier Ferret. ń Approche endogène et exogène pour améliorer la segmentation thématique de documents ż. *TAL*, 2006.

[Ferret *et al.*, 2001] Olivier Ferret, Brigitte Grau, Jean-Luc Minel, et Sylvie Porhiel. ń Repérage de structures thématiques dans des textes ż. *TALN2001*, 2001.

[Garner, 1995] Stephen R. Garner. ń WEKA : The Waikato Environment for Knowledge Analysis ż. pp 57–64, 1995.

[Georgescul *et al.*, 2006] Maria Georgescul, Alexander Clark, et Armstrong Susan. ń Word Distributions for Thematic Segmentation in a Support Vector Machine Approach ż. 2006.

[Grosz & Sidner, 2002] Barbara J. Grosz et Candace L. Sidner. ń Attention, Intentions, And The Structure Of Discourse ż. 2002, unknown.

[Hearst, 1993] Marti A. Hearst. ń TextTiling : A quantitative approach to discourse segmentation ż. , 1993.

[Hearst, 1994] M. A. Hearst. ń Multi-paragraph segmentation of expository text ż. *32th Annual Meeting of the Association for Computational Linguistics*, pp 9–16, 1994.

[Hearst, 1997] M. A. Hearst. ń TextTiling : Segmenting text into multi-paragraph subtopic passages. ż. *Computational Linguistics.*, pp 33–64, 1997.

[Helfman, 1994] J. Helfman. ń Similarity Patterns in Language ż. *Visual Languages*, pp 173–175, 1994.

[HP2006] Martine Hurault-Plantet, Michèle Jardino, et Jean-Baptiste Berthelin. ń Ajustement des frontières de segments thématiques détectés automatiquement ż. 2006.

[Jakobson, 1963] Roman Jakobson. *Essai de linguistique générale*. 1963.

[Ji & Zha, 2003] X. Ji et H. Zha. ń Domain-independant segmentation using anisotropic diffusion and dynamic programming ż. *Proceedings of ACM/SIGIR Conference of Research and Developpement in Information Retrieval*, 2003.

[Kan *et al.*, 1998] M. Kan, J. L. Klavans, et K. R. McKeown. ń Linear Segmentation and Segment Significance ż. *Proceedings of WVLC-6*, pp 197–205, 1998.

[Karatzas, 2003] D. Karatzas. *Text Segmentation in Web Images Using Color Perception and Topological Features*. ECS Publications, UK, 2003.

[Kaufmann, 1999] Stefan Kaufmann. ń Cohesion and collocation : using context vectors in text segmentation ż. Dans les actes de *Proceedings of the 37th annual meeting of the ACL*, pp 591–595, 1999.

[Kozima, 1993] Hideki KOZIMA. ń Text segmentation based on similarity between words ż. *31th annual meeting of the Association for Computational Linguistics*, pp 286–288, 1993.

[Labadié & Chauché, 2006] A. LABADIÉ et J. CHAUCHÉ. ń Segmentation thématique par calcul de distance sémantique ż. *Actes de DEFT'06*, pp 45–59, 2006.

[Labadié & Prince, 2008a] A. LABADIÉ et V. PRINCE. ń Comparaison de méthodes lexicales et syntaxico-sémantiques dans la segmentation thématique de texte non supervisée ż. *Actes de TALN'08*, 2008.

[Labadié & Prince, 2008b] A. LABADIÉ et V. PRINCE. ń Finding text boundaries and finding topic boundaries : two different task ? ż. *Proceedings of GoTAL 2008*, 2008.

[Labadié & Prince, 2008c] A. LABADIÉ et V. PRINCE. ń Intended boundaries detection in topic change tracking for text segmentation ż. *Proceeding of NLPCS 2008*, 2008.

[Labadié & Prince, 2008d] Alexandre LABADIÉ et Violaine PRINCE. ń The impact of corpus quality and type on topic based text segmentation evaluation ż. Dans les actes de *CLA 2008*, 2008.

[Lafourcade & Prince, 2001] M. LAFOURCADE et V. PRINCE. ń Synonymie et vecteurs conceptuels ż. *Proceedings of TALN'01*, pp 233–242, 2001.

[Lamprier *et al.*, 2007] Sylvain LAMPRIER, Tassadit AMGHAR, Bernard LEVRAT, et Frederic SAUBION. ń SegGen : a Genetic Algorithm for Linear Text Segmentation ż. *Proceedings of IJCAI'07*, 2007.

[Lamprier *et al.*, 2008] Sylvain LAMPRIER, Tassadit AMGHAR, Bernard LEVRAT, et Frederic SAUBION. ń Using an Evolving Thematic Clustering in a Text Segmentation Process ż. *Universal Computer Science*, pp 178–192, 2008.

[Landauer & Dumais, 1997] T. LANDAUER et S. DUMAIS. ń A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge ż. *Psychological Review*, pp 211–240, 1997.

[Landauer *et al.*, 1998] T. LANDAUER, P. FOLLTZ, et D. LAHAM. ń An introduction to Latent Semantic Analysis ż. *Discourse Processes*, pp 259–284, 1998.

[Larousse, 1992a] LAROUSSE. *Dictionnaire Larousse*. Larousse, 1992.

[Larousse, 1992b] Larousse. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, Paris, 1992.

[Larousse, 1997] Larousse, . *Petit Larousse Illustré*. Larousse, 1997.

[Lelu *et al.*, 2006] A. Lelu, Cadot M., et S. Aubain. ń Coopération multiniveau d'approches non-supervisées et supervisées pour la detection des ruptures thématiques dans les discours présidentiels français ż. *Actes de DEFT'06*, 2006.

[Lin & Hovy, 1997] Chin-Yu Lin et E. Hovy. ń Identifying topics by position ż. Dans les actes de *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, pp 283–290., 1997.

[Mann & Thompson, 1987] W. Mann et S. A. Thompson. ń Rhetorical Structure Theory : A Theory of Text Organization. ż. ISI/RS-87-190, University of Southern California, Information Sciences Institute, 1987.

[McDonald & Chen, 2002] D McDonald et H. Chen. ń Using sentence selection heuristics to rank text segments in txtractor ż. Dans les actes de *Proceeding of the 2nd ACM/IEEE Joint Conference on Digital Librairies*, pp 28–35, 2002.

[Morris & Hirst, 1991] J. Morris et G. Hirst. ń Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text ż. *Computational Linguistics*, pp 20–48, 1991.

[Palmer & Hearst, 1997] David D. Palmer et Marti A. Hearst. ń Adaptive multilingual sentence boundary disambiguation ż. *Computational Linguistics*, pp 241–267, 1997.

[Passonneau & Litman, 1993] R. J. Passonneau et D.J. Litman. ń Intention-based segmentation : Humanreliability and correlation with linguistic cues ż. *Proceedings of the 31st Annual Meeting of theAssociation for Computational Linguistics,*, pp 148–155, 1993.

[Passonneau & Litman, 1997] Rebecca J. Passonneau et Diane J. Litman. ń Discourse Segmentation by Human and Automated Means ż. *Computational Linguistics*, pp 103–139, 1997.

[Pelleg & Moore, 2000] D. Pelleg et A. W. Moore. ń X-means : Extending K-means with Efficient Estimation of the Number of Clusters ż. Dans les actes de *Seventeenth International Conference on Machine Learning*, pp 727–734. Morgan Kaufmann, 2000.

[Pevzner & Hearst, 2002] Lev Pevzner et Marti Hearst. ń A Critique and Improvement of anEvaluation Metric for Text Segmentation ż. *Computational Linguistics*, pp 113–125, 2002.

[Prince & Labadié, 2007] V. Prince et A. Labadié. ń Text Segmentation Based on Document Understanding for Information Retrieval. ż. *Proceedings of NLDB'07*, pp 295–304, 2007.

[Reynar, 1998] Jeffrey C. Reynar. ń Topic Segmentation : Algorithms and Applications. ż. *Phd thesis, University of Pennsylvania*, 1998.

[Reynar, 2002] Jeffrey C. Reynar. ń Statistical Models for Topic Segmentation ż. 2002, unknown.

[Roget, 1852] P. Roget. *Thesaurus of English Words and Phrases*. Longman, London, 1852.

[Salton & Lesk, 1965] Gerard Salton et M. E. Lesk. ń The SMART automatic document retrieval systemsŮan illustration ż. *Communications of the ACM*, pp 391–398, 1965.

[Salton *et al.*, 1975] Gerard Salton, A. Wong, et C. S. Yang. ń A Vector Space Model for Automatic Indexing ż. *Commun. ACM*, pp 613–620, 1975.

[Schmid, 1994] Helmut Schmid. ń Probabilistic Part-of-Speech Tagging Using Decision Trees ż. 1994.

[Sitbon & Bellot, 2004] Laurianne Sitbon et Patrice Bellot. ń Evaluation de méthodes de segmentation thématique linéaire non supervisés aprés adaptation au français ż. *Actes de TALN'04*, 2004.

[Sitbon, 2004] Laurianne Sitbon. ń Fusion d'approches non supervisées et génériques pour la segmentation thématique ż. 2004.

[Stokes *et al.*, 2004] Nicola Stokes, Joe Carthy, et Alan F. Smeaton. ń SeLeCT : A Lexical Cohesion Based News Story Segmentation System ż. 2004.

[Utiyama & Isahara, 2001] M. Utiyama et H Isahara. ń A statistical model for domain-independent text segmentation ż. *ACL*, pp 491–498, 2001.

[Wu & Tseng, 1993] Z. Wu et G. Tseng. ń Chinese Text Segmentation for Text Retrieval : Achievements and Problems. ż. *Journal of the American Society for Information Science*, pp 532–542, 1993.

[Yang & Li, 2005] C. C. Yang et K. W. Li. ń A heuristic method based on a statistical approach for Chinese text segmentation. ż. *Journal*

of the American Society for Information Science and Technology, pp 1438–1447, 2005.

[YM2007] Mehdi Yousfi-Monod. ń *Compression automatique ou semi-automatique de texte par élagage des constituants effaçable : une approche interactive et indépendante des corpus.* ż. PhD thesis, Université des Sciences et technique du Languedoc, 2007.

Ce travail s'inscrit dans le domaine du traitement automatique du langage naturel et traite plus spécifiquement de l'application de ce dernier à la segmentation thématique de texte. L'originalité de cette thèse consiste à intégrer dans une méthode non-supervisée de segmentation thématique de texte de l'information syntaxique, sémantique et stylistique. Ce travail propose une approche linéaire de la segmentation thématique s'appuyant sur une représentation vectorielle issue de l'analyse morpho-syntaxique et sémantique de la phrase. Cette représentation est ensuite utilisée pour calculer des distances entre segments thématiques potentiels en intégrant de l'information stylistique. Ce travail a donné lieu au développement d'une application qui permet de tester les différents paramètre de notre modèle, mais qui propose également d'autres approches testées dans ce travail. Notre modèle a été évaluer de deux manière différente, une évaluation automatique sur la base de textes annotés et une évaluation manuelle. Notre évaluation manuelle a donné lieu à la définition d'un protocole d'évaluation s'appuyant sur des critères précis. Dans les deux cas, les résultats de notre évaluation ont été au niveau, voir même au dessus, des performances des algorithmes les plus populaires de la littérature.

This research belongs to the Natural Language Processing (NLP) field and more specifically focuses on topic text segmentation. The originality of this thesis consists in integrating to an unsupervised topic text segmentation method syntactic, semantic and stylistic information. This work present a linear approach of topic text segmentation based on a vectorial representation of the sentence coming from a deep morpho-syntactic and semantic analysis. This representation is then used to compute distance between potential topic segment while integrating stylistic information. During this research an application has been developed, this application allows users to test the approach various parameters, but also some others methods that have been tested during this research. Our model has been evaluated using an automatic evaluation and a manual evaluation. Our manual evaluation leas us to develop a specific evaluation protocol for the task based on precise parameters. In both automatic and manual evaluation our results as good and sometimes even better than some of the most popular algorithms.

Discipline : Informatique

Laboratoire : Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier (LIRMM) ; UMR 5506 ; 161 rue Ada, 34392 Montpellier Cedex 5, France