Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole

Gwénolé Lecorvé

Thèse

THÈSE INSA Rennes
sous le sceau de l'Université européenne de Bretagne
pour obtenir le grade de
DOCTEUR DE L'INSA DE RENNES
Spécialité : Informatique

présentée par
Gwénolé LECORVÉ
ÉCOLE DOCTORALE : MATISSE
LABORATOIRE : *IRISA/INRIA (équipe TexMex)*

Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole

Thèse soutenue le 24.11.2010
devant le jury composé de :

Frédéric BÉCHET
Professeur à l'Université de la Méditerranée, LIF / Président
Yannick ESTÈVE
Professeur à l'Université du Maine, LIUM / Rapporteur
Marie-Francine MOENS
Professeur à la Katholieke Universiteit Leuven / Rapporteur
Gilles ADDA
Ingénieur de recherche CNRS, LIMSI / Examinateur
Pascale SÉBILLOT
Professeur à l'INSA de Rennes, IRISA / Directrice de thèse
Guillaume GRAVIER
Chargé de recherche CNRS, HDR, IRISA / Co-directeur de thèse

Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole

Gwénolé LECORVÉ

En partenariat avec

Si on fait l'amour, c'est pour satisfaire les sens. Et c'est pour l'essence qu'on fait la guerre ! D'ailleurs, la plupart des gens préfèrent glisser leur peau sous les draps que de la risquer sous les drapeaux.

Raymond Devos

Remerciements

En dehors des aspects incontournables d'une thèse que sont la réflexion scientifique, la mise en place d'expériences et la communication de résultats, le doctorat a également été pour moi un chouette moment de vie. Je tiens à remercier ceux qui ont pris part à mes côtés à cette aventure unique.

Tout d'abord, je tiens à remercier Frédéric Béchet de m'avoir fait l'honneur de présider mon jury de thèse. J'adresse également des remerciements sincères à Yannick Estève et Marie-Francine Moens pour avoir accepté de relire mon travail et d'y avoir porté un regard critique. Je remercie aussi Gilles Adda, bien-sûr pour sa participation à mon jury, mais aussi pour son amabilité infaillible ainsi que la richesse et la pertinence de ses remarques.

Par ailleurs, je ne peux résumer ces quelques années de travail sans évoquer l'ambiance formidable de l'équipe TexMex. Je souhaite à tout doctorant de pouvoir effectuer sa thèse dans de si bonnes conditions. Je tiens ainsi à remercier chaque membre de l'équipe pour m'avoir aussi bien accueilli et intégré. J'adresse des remerciements tout particuliers aux membres du groupuscule RAPTAL (Fabienne, Christian, Julien...) et aux adeptes des déjeuners à la cafét' (Vincent, Laurent, Ewa...) auprès desquels les discussions, qu'elles aient concerné le travail ou tout autre chose, furent toujours l'occasion de moments plaisants. Plus spécifiquement, je remercie Patrick qui, malgré ses nombreuses responsabilités, a toujours été accessible et ouvert à l'écoute, ainsi que Camille avec qui j'ai eu le plaisir de partager le bureau pendant deux ans dans une ambiance, certes, studieuse mais avant tout chaleureuse.

Au delà de l'équipe, je tiens également à remercier différents autres collègues. Je remercie ainsi les éminents membres actifs de l'ADOC et de Nicomaque avec qui j'ai pris beaucoup de plaisir à sortir du cadre pur de la recherche pour m'ouvrir à d'autres considérations. Je pense aussi bien entendu à Cédric. Quel chance d'avoir pu passer cette thèse, mais également toutes mes études, avec un si bon ami. Sa complicité quasi quotidienne m'a apporté beaucoup de soutien. Je lui exprime tous mes vœux de réussite pour la fin de ses travaux de thèse.

J'aimerais également remercier tous ceux qui, en dehors de l'Irisa, ont soit pris le temps de s'intéresser à mon travail ou ont participé à m'aérer l'esprit dans les moments difficiles.

Pour ne citer qu'eux, merci donc à Erwan, Michel, Nico, Pierre-Marie, Thomas et Typhène. J'adresse également de très vifs remerciements à Marie qui a toujours été là pour m'écouter, me soutenir, m'encourager mais aussi pour partager mes réussites. Assurément, l'aventure n'aurait pas été aussi belle sans elle. Enfin, je tiens de tout mon cœur à remercier ma famille, et plus particulièrement mes parents, qui ont toujours cru en moi et approuvé mes divers choix d'orientation. Je leur dédie ce travail car c'est grâce à eux que j'en suis là aujourd'hui.

Pour conclure – *last but not least* comme on a l'habitude de dire – , je tiens à adresser toute ma gratitude à Pascale et Guillaume pour m'avoir fait le grand plaisir de m'accepter comme stagiaire de master puis comme doctorant. Ils sont un exemple de compétence, de rigueur scientifique, de disponibilité et de gentillesse. Merci à eux de m'avoir insufflé leur passion pour la recherche. J'ai notamment pris beaucoup de plaisir à assister à leur côté à l'émergence de nouvelles problématiques nées de la convergence, déjà entamée avec la thèse de Stéphane Huet, de leurs disciplines respectives d'origine que sont le traitement automatique des langues et la reconnaissance automatique de la parole. Je leur souhaite bonne continuation dans cette captivante aventure.

Table des matières

Introduction

Contexte d'étude

En tant que mode de communication privilégié de l'humain, la parole aiguise depuis longtemps la curiosité des scientifiques. Que ce soit en vue d'une meilleure compréhension du processus de pensée humain ou à des fins technologiques telles que le dialogue homme-machine, beaucoup de travaux ont cherché à formaliser la parole afin d'en permettre la reconnaissance, c'est-à-dire de permettre le passage automatique d'un signal de parole à du texte. Avec l'avènement de l'informatique, ces efforts ont été récompensés par de nombreux progrès. Aujourd'hui, les systèmes les plus polyvalents, systèmes dits de reconnaissance de la parole continue à grand vocabulaire, commencent à transcrire raisonnablement bien des discours prononcés par des locuteurs variés avec un vocabulaire riche. Ces systèmes ouvrent la voie à de nouvelles applications, notamment dans le domaine du traitement de documents multimédias.

Avec la popularisation récente de dispositifs d'enregistrement numérique ainsi que l'essor de nouvelles technologies de télécommunication, la production et la consommation de contenus multimédias ont explosé. À titre d'exemple, plus de 150 chaînes sont disponibles par défaut à partir des boîtiers TV fournis par de nombreux fournisseurs d'accès à Internet. De même sur Internet, le site de partage de vidéos YouTube annonçait début 2010 référencer chaque minute l'équivalent de 24 heures de nouvelles vidéos, tout en revendiquant 2 milliards de vidéos vues par jour. Face à ce foisonnement, l'automatisation de tâches facilitant l'accès au contenu de documents multimédias est devenue un enjeu majeur. Citons parmi ces tâches l'indexation de documents afin de faciliter les recherches d'utilisateurs, la construction automatique de résumés pour permettre une compréhension rapide des documents, ou encore la mise en relation de documents multimédias pour favoriser les possibilités de navigation entre contenus. De manière générale, ces tâches nécessitent d'accéder automatiquement au contenu sémantique des documents. Or, parmi les différentes modalités offertes par le multimédia (vidéo, image, son, texte...), la parole s'avère être un vecteur d'information très riche dans une majorité de cas. En cela, les systèmes de reconnaissance automatique de la parole sont des outils primordiaux car, en transcrivant un signal de parole en texte, ils ouvrent la voie à des traitements élaborés à partir de techniques de traitement automatique des langues. Dans cette optique, la qualité des transcriptions fournies par un système, typiquement le nombre de mots mal transcrits, est un paramètre essentiel. Il est notamment primordial de bien transcrire les mots ou séquences de mots les plus

porteurs de sens dans un document si l'on souhaite en décrire correctement le contenu.

Dans ce contexte, on attend alors des systèmes de reconnaissance automatique de la parole qu'ils soient capables de transcrire efficacement une très large panoplie de documents multimédias se distinguant notamment par des sujets abordés très variés. Pour cela, une grande majorité de systèmes actuels s'appuie sur un cadre statistique permettant de concilier les aspects acoustiques et linguistiques du langage parlé. Sur le plan linguistique, le langage s'y trouve modélisé par un vocabulaire, qui recense l'ensemble figé des mots que sera capable de reconnaître le système, et par un modèle de langue qui a pour but d'assigner une probabilité *a priori* à des séquences de mots du vocabulaire. Pour calculer ces probabilités, les modèles de l'état de l'art ne modélisent pas des séquences de mots dans leur globalité mais calculent localement la probabilité de chacun de leurs mots étant donnés quelques mots qui les précèdent. Ces modèles, appelé *modèles n-grammes*, sont ceux qui permettent aujourd'hui d'obtenir les meilleures performances pour des tâches de reconnaissance automatique de la parole à grand vocabulaire. Cette modélisation statistique du langage requiert un processus d'apprentissage relativement simple puisqu'il ne nécessite aucune expertise linguistique. Celui-ci s'appuie en effet principalement sur le calcul de fréquences d'apparition de séquences de mots au sein d'un très vaste corpus textuel d'apprentissage. Cependant, malgré cette simplicité apparente, cet apprentissage nécessite en réalité une multitude de petits traitements fastidieux, notamment liés à la préparation des textes. Par conséquent, une fois le vocabulaire et le modèle de langue appris, il est exclu de lancer un nouvel apprentissage pour essayer de modéliser au mieux le contenu linguistique de chaque nouveau document que le système de reconnaissance aurait à transcrire. Pourtant, il ressort de manière évidente que ce contenu varie grandement selon que les documents abordent un thème ou un autre. Par exemple, le choix et l'emploi des mots faits par un locuteur n'est pas le même s'il parle de ses vacances dans le Pacifique ou s'il parle du dernier accessoire *high-tech* qui vient de sortir. Pour compenser cette impossibilité, le corpus d'apprentissage utilisé est souvent constitué de textes d'une grande variété thématique afin de fournir au vocabulaire et au modèle de langue une connaissance la plus vaste possible des usages de la langue. Si cette stratégie fonctionne relativement bien dans le cadre de la transcription d'une collection fermée de documents dont les thèmes abordés sont relativement proches de ceux vus dans le corpus d'apprentissage, elle atteint ses limites dans le cadre du traitement de flux multimédias ou de collections fluctuantes telles que les bases de sites de partage de vidéos. Dans ces dernières conditions, les thèmes abordés changent fréquemment et sont difficilement prévisibles. Il devient alors irréaliste d'espérer pouvoir construire un corpus d'apprentissage qui couvre tous les thèmes envisageables une fois pour toutes et qui permette une utilisation pertinente pour chacun des thèmes. Pour pallier ce problème, une solution est d'adapter le vocabulaire et le modèle de langue au thème de chaque nouveau document rencontré sans relancer un processus complet d'apprentissage. C'est précisément ce à quoi s'intéresse cette thèse.

Problématique et objectifs de la thèse

Cette thèse présente nos travaux sur l'adaptation thématique d'un système de reconnaissance automatique de la parole. Pour réaliser cette adaptation, nous adoptons une stratégie en 3 temps présentée par la figure 1. Dans un premier temps, partant d'un document à transcrire supposé traiter d'un thème unique mais non connu à l'avance, nous

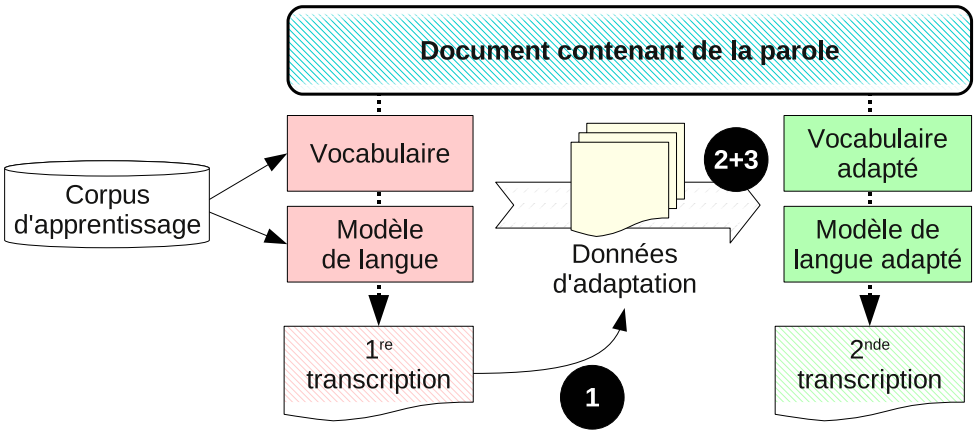


FIGURE 1 – Stratégie d'adaptation thématique d'un système de reconnaissance automatique de la parole pour un document multimédia donné.

cherchons à récupérer des données textuelles relevant du même thème que le document en nous appuyant sur une première transcription obtenue grâce au système de reconnaissance généraliste à adapter. En se fondant sur les informations que fournissent ces données, il s'agit alors, dans un second temps, d'enrichir le vocabulaire du système afin de couvrir au mieux la richesse lexicale du thème considéré et, dans un troisième temps, de mettre à jour le modèle de langue de manière à ce que ce dernier modélise mieux les probabilités de succession des mots au sein du thème. L'objectif final de ce processus est de générer une seconde transcription du document initial avec l'espoir que celle-ci soit de meilleure qualité que celle fournie par le système généraliste. Elle doit notamment contenir moins d'erreurs sur les séquences de mots spécifiques au thème du document de manière à permettre des traitements sémantiques performants en aval de la tâche de transcription, comme par exemple une tâche d'indexation.

Ce schéma d'adaptation correspond globalement à une tendance déjà amorcée dans divers travaux d'adaptation linguistique d'un système de transcription (Bellegarda, 2004). Il ressort de ces études que la mise en œuvre d'un tel processus doit résoudre de multiples problèmes. Tout d'abord, la manière de caractériser un thème à partir d'une transcription automatique est une question centrale car cette caractérisation doit guider la tâche de récupération de données d'adaptation et conditionne le reste du processus. De même, il s'avère important de savoir déterminer ce qui différencie l'emploi usuel du langage – nous parlerons de langage *général* ou *généraliste* – de son emploi au sein d'un thème donné. Par ailleurs, il faut réfléchir au choix des ressources textuelles qui permettront de dévoiler ces spécificités. De manière complémentaire, il faut être à même de pouvoir intégrer ces spécificités dans un nouveau vocabulaire et dans un nouveau modèle de langue. Ceci soulève des questions plus techniques concernant les méthodes permettant de modifier les probabilités d'un modèle de langue, d'ajouter de nouveaux mots dans le vocabulaire et d'intégrer ces derniers dans le nouveau modèle de langue.

Dans notre travail, la réalisation de ces diverses tâches est guidée par une même philosophie qui distingue notre travail des tentatives rencontrées jusqu'à présent dans la littérature. Tout d'abord, nous avons développé une chaîne d'adaptation complète qui, en tant que telle, se démarque d'une majorité de travaux qui n'étudient qu'un aspect du schéma d'adaptation que nous avons présenté. Par ailleurs, nous avons souhaité que l'ensemble des

traitements de cette chaîne soit entièrement non supervisé de manière à pouvoir traiter des documents tout-venant de manière automatique. Cette volonté concerne notamment les méthodes que nous souhaitons adopter pour la caractérisation d'un thème. Nous ne voulons en particulier pas faire quelque hypothèse que ce soit sur les thèmes que le système de transcription pourrait être amené à rencontrer et, à plus forte raison, nous ne voulons pas nous fonder sur un ensemble figé de thèmes fixés *a priori* car nous ne savons pas à l'avance quels thèmes vont être abordés et ceux-ci peuvent évoluer avec le temps. Par ailleurs, nous ne souhaitons pas figer la caractérisation d'un thème car celle-ci peut s'avérer différente d'un document à l'autre. Pour combler cette absence de supervision, nous faisons un usage intensif de techniques de traitement automatique des langues.

Concrètement, nous proposons de caractériser le thème d'une transcription par des mots-clés extraits automatiquement en adaptant le critère *tf-idf*, issu du domaine de la recherche d'information et classiquement appliqué à des textes écrits, aux spécificités de l'oral. Nous montrons que cette caractérisation d'un thème permet de filtrer thématiquement des pages Web afin de constituer un corpus d'adaptation. Par ailleurs, nous mettons au jour les spécificités linguistiques de ce corpus en nous appuyant sur des techniques d'acquisition automatique de terminologies, techniques qui consistent à mettre en évidence les mots et séquences de mots importants pour un thème donné. Ces deux techniques présentent l'originalité et l'intérêt de manipuler des éléments linguistiques explicites, tels que des mots ou des séquences de mots, qui sont directement intégrables dans les composants linguistiques d'un système de reconnaissance automatique de la parole. Elles se distinguent en cela particulièrement des méthodes dérivées de l'analyse sémantique latente qui s'appuient sur des concepts flous et difficilement utilisables tels quels.

Enfin, nous souhaitons rester dans le cadre de la modélisation du langage par des modèles n-grammes car ceux-ci sont les modèles les plus performants actuellement et sont en cela ceux qu'utilise la plupart des systèmes récents de reconnaissance automatique de la parole à grand vocabulaire[1]. Dans ce cadre, nous proposons deux contributions originales. D'une part, nous montrons que le processus de réestimation des probabilités d'un modèle de langue peut être rendu fonction de chaque thème sans aucune supervision. Pour cela, nous spécialisons la technique dite d'adaptation par minimum d'information discriminante (MDI) de telle sorte que seules soient adaptées les probabilités du modèle de langue portant sur les mots importants pour le thème considéré, ces mots étant fournis par nos terminologies. Cette méthode se distingue particulièrement des techniques de réestimation classiques qui n'intègrent aucune connaissance linguistique. D'autre part, nous présentons une nouvelle manière d'intégrer de nouveaux mots liés à un thème donné dans le vocabulaire et le modèle de langue d'un système initial. Nous montrons notamment qu'il est possible de déterminer automatiquement les séquences de mots qui se rapportent le plus vraisemblablement à de nouveaux mots, alors que ceux-ci n'ont jamais ou que rarement été observés lors de l'apprentissage du système. En assimilant chaque nouveau mot à des mots déjà connus du système grâce à des relations paradigmatiques, cette technique présente l'avantage de générer des n-grammes sémantiquement cohérents et ainsi de ne pas surgénéraliser l'emploi d'un nouveau mot comme le font généralement des modèles de langue n-grammes à vocabulaire ouvert.

[1] Nous restons néanmoins conscient que d'autres types de modélisations linguistiques pourraient être intéressants pour mettre en application un processus d'adaptation thématique. Nous ne manquerons pas de souligner diverses pistes qui pourraient, à l'occasion, être investiguées en ce sens.

Organisation du manuscrit

Ce manuscrit se compose de sept chapitres. Les trois premiers posent le cadre scientifique de nos travaux. Nous présentons tout d'abord une vue d'ensemble du fonctionnement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire tel que celui que nous avons utilisé pour nos travaux. Nous détaillons ensuite la question de la modélisation statistique du langage, dressant un panorama des différentes approches théoriques existantes et nous attardant sur le cas des modèles n-grammes. Ce chapitre est pour nous l'occasion de développer les faiblesses de ces modèles, notamment pour modéliser le langage dans un contexte thématiquement marqué. Le chapitre 3 présente alors un état de l'art des travaux qui s'intéressent à l'adaptation des composants linguistiques d'un système de reconnaissance automatique de la parole. À la lumière de ces explications et des enseignements qui peuvent en être tirés, le chapitre 4 présente une vue détaillée de la chaîne de traitements que nous proposons. Comme nous l'avons expliqué, cette chaîne s'articule autour des tâches de récupération de données d'adaptation, de réestimation des probabilités du modèle de langue et de mise à jour du vocabulaire. Les chapitres suivants détaillent alors nos contributions pour chacune de ces tâches. Au chapitre 5, nous présentons une technique de récupération non supervisée de textes thématiques à partir d'Internet. Cette technique s'appuie sur une caractérisation d'un thème par des mots-clés extraits d'une transcription automatique. Le chapitre 6 montre ensuite comment il est possible de caractériser les spécificités lexicales et linguistiques d'un thème à partir des données d'adaptation par la construction de terminologies. Nous montrons que ces terminologies peuvent permettre d'adapter précisément un modèle de langue n-gramme. Enfin, le chapitre 7 présente des travaux plus exploratoires sur la mise à jour du vocabulaire. Nous mettons notamment en lumière le fait que des mots jusqu'alors absents du vocabulaire peuvent être intégrer dans un modèle de langue n-gramme pré-existant en générant artificiellement des contextes d'utilisation de ces nouveaux mots grâce à des relations sémantiques.

Chapitre 1

Reconnaissance automatique de la parole

Dans notre perspective d'adaptation thématique des composants linguistiques d'un système de reconnaissance automatique de la parole, il est essentiel d'avoir une bonne vision d'ensemble du processus permettant de passer d'un signal audio à sa transcription textuelle. L'objectif de ce chapitre est donc de dresser un panorama tant théorique que pratique du fonctionnement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire. Plus particulièrement, nous mettons l'accent sur les différents points du processus de transcription qui présentent des aspects linguistiques. Pour plus de précisions, signalons qu'il existe différents très bons ouvrages qui exposent en détails le fonctionnement de tous les aspects d'un système de transcription (Rabiner et Juang, 1993 ; Jelinek, 1998 ; Haton *et al.*, 2006). Dans un premier temps, nous présentons les principes généraux de la reconnaissance statistique de la parole, puis montrons par quels biais algorithmiques ces principes peuvent être efficacement mis en œuvre. Nous nous attardons ensuite sur quelques types de sorties que peut fournir un système de reconnaissance ainsi que sur différentes méthodes d'évaluation existantes pour juger de leur qualité. Enfin, nous donnons un exemple concret de système à travers la présentation du système IRENE, système utilisé pour l'ensemble des travaux relatés dans ce manuscrit.

1.1 Principes généraux

Le but d'un système de reconnaissance automatique de la parole est de fournir la transcription textuelle d'un signal audio d'entrée contenant de la parole. Dans le cadre de la modélisation statistique de la parole, cette tâche équivaut à rechercher parmi l'ensemble des séquences de mots possibles à partir d'un vocabulaire fixé la séquence la plus probable W^* étant donnée une séquence Y de caractéristiques acoustiques observées à partir du signal d'entrée. Mathématiquement, cela s'écrit sous la forme de la maximisation *a posteriori* suivante :

$$W^* = \arg\max_W P[W|Y] \ . \tag{1.1}$$

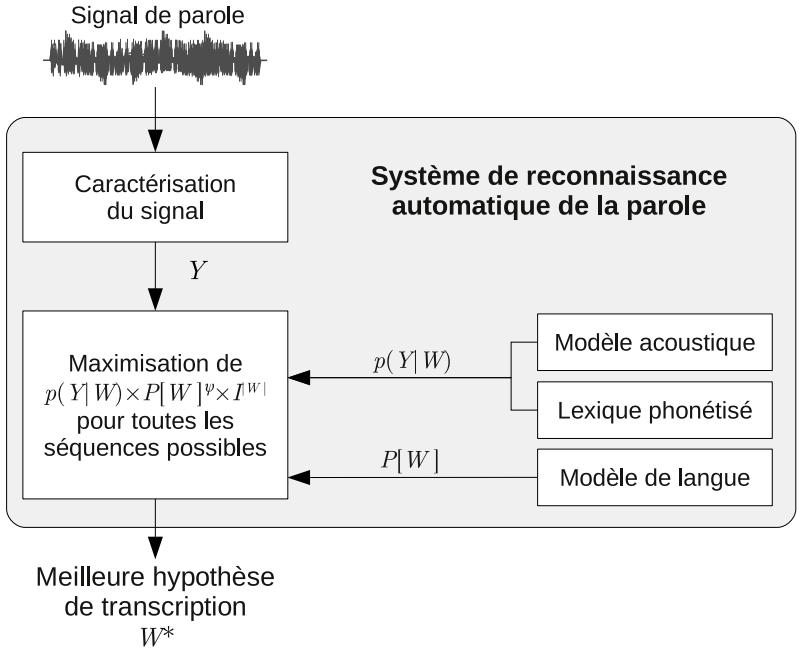


FIGURE 1.1 – Vue schématique d'un système de reconnaissance automatique de la parole.

où W parcourt l'ensemble des séquences de mots possibles à partir du vocabulaire du système. Par application de la règle de Bayes, cette formule est décomposée en :

$$W^* = \arg\max_W \frac{p(Y|W) \times P[W]}{p(Y)} \tag{1.2}$$

où $p(Y|W)$ est la vraisemblance des observations acoustiques sachant une séquence de mots testée, $P[W]$ est la probabilité *a priori* de cette séquence de mots et $p(Y)$ est la vraisemblance *a priori* de la réalisation acoustique. Puisque la vraisemblance $p(Y)$ est la même quelque soit la séquence W, la recherche de W^* peut se simplifier en :

$$W^* = \arg\max_W p(Y|W) \times P[W] \ . \tag{1.3}$$

En pratique, la formule (1.3) subit quelques aménagements. D'une part, les scores $p(Y|W)$ et $P[W]$ ne sont pas du même ordre de grandeur car ils ne sont pas calculés sur le même type de données. Par conséquent, un facteur d'échelle ψ est introduit pour pondérer les probabilités linguistiques $P[W]$. D'autre part, il s'avère nécessaire d'ajouter un terme permettant de jouer sur la longueur des séquences de mots souhaitée en sortie. Ce coefficient, noté I, est appelé *pénalité d'insertion*. La recherche de la séquence optimale W^* s'appuie alors sur le calcul d'un score global pour chaque hypothèse de transcription W, soit :

$$W^* = \arg\max_W \underbrace{p(Y|W) \times P[W]^{\psi} \times I^{|W|}}_{score(W)} , \tag{1.4}$$

où $|W|$ est le nombre de mots de la séquence W.

Cette dernière formule met en évidence différents composants (figure 1.1) : un module de caractérisation du signal permet de transformer le signal audio en une séquence Y de

Orthographe → Prononciation
clans → k l a$^{\sim}$
clans → k l a$^{\sim}$ z
clef → k l e
clé → k l e
être → E t R @
être → E t R
être → E t

TABLE 1.1 – Aperçu du contenu du vocabulaire.

vecteurs numériques ; un *modèle acoustique* calcule la vraisemblance $p(Y|W)$; un *modèle de langue* se charge de la probabilité $P[W]$, alors que le lien entre les représentations sur lesquelles se fondent ces deux derniers modèles se fait par l'entremise d'un *lexique phonétisé* qui associe à chaque mot du vocabulaire une ou plusieurs prononciations possibles. Nous décrivons brièvement chacun de ces quatre composants.

1.1.1 Caractérisation du signal

La représentation adoptée d'un signal de parole consiste en une séquence de vecteurs numériques $Y = y_1...y_T$ où chaque vecteur y_i représente quelques millisecondes (typiquement 10 ms) du signal d'entrée. De ces « tranches » de signal, appelées *trames*, des caractéristiques relatives à l'énergie et aux gammes de fréquences vocales sont extraites. En incluant les variations du premier, voire du second ordre, de ces paramètres, les vecteurs de caractéristiques sont typiquement de dimension 40.

1.1.2 Vocabulaire et lexique phonétisé

Le vocabulaire définit l'ensemble des mots qu'est capable de manipuler le système de reconnaissance. En cela, cet ensemble est un élément déterminant car il restreint les sorties du système aux seuls mots qu'il contient. Typiquement, dans un système dit « à grand vocabulaire », le vocabulaire répertorie plusieurs dizaines de milliers de mots. En pratique, pour faire le lien entre modélisations acoustique et linguistique, ce vocabulaire est transformé en un lexique phonétisé (ou dictionnaire de prononciation, ou dictionnaire phonétique) où chaque mot est associé à une liste de prononciations possibles. Ces prononciations sont représentées sous la forme de séquences de phonèmes, unités représentant les sons élémentaires d'une langue et utilisant un alphabet propre.

La table 1.1 donne un aperçu du contenu potentiel d'un lexique phonétisé. Cet exemple nous offre l'occasion de quelques précisions. Tout d'abord, les mots du vocabulaire ne sont symbolisés que par une chaîne graphique représentant leur orthographe, c'est-à-dire qu'aucune information sémantique ne leur est attachée. Ainsi, du point de vue du système de reconnaissance, deux mots seront acoustiquement et linguistiquement modélisés de la même manière s'ils partagent la même orthographe, quand bien même cette orthographe puisse confondre différentes acceptions. Par exemple, le mot **être** peut référer à l'auxiliaire tout comme au nom commun. À l'inverse, les mots **clé** et **clef** seront considérés comme deux mots distincts. Ces problèmes peuvent être levés au moment de la construction du vocabulaire ou par l'adoption de modèles de langues particuliers comme nous le verrons au

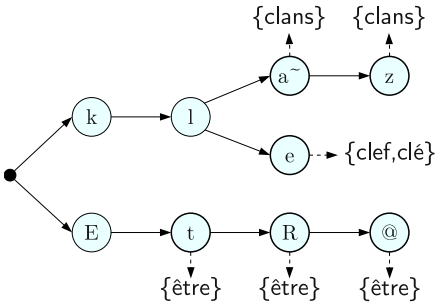


FIGURE 1.2 – Représentation du lexique phonétisé sous la forme d'un arbre lexical. Cerclés de gras, les nœuds correspondant à la fin de transcription phonétique d'un mot.

chapitre 2. Ensuite, au sein du lexique phonétisé, les variantes de prononciation d'un mot peuvent permettre de modéliser les prononciations résultant d'éventuelles liaisons avec les mots qui pourraient les précéder ou leur succéder. Comme la multiplication des variantes de prononciation dans un lexique tend toutefois à ajouter de l'ambiguïté lors du processus de transcription (Rosenfeld, 1995), certains systèmes proposent de contextualiser certaines d'entre elles. Par exemple, il est possible de préciser que la variante de prononciation « clans → k l a~ z », où le **s** final est prononcé, n'est valable que dans un contexte où **clans** est suivi d'un mot commençant par une voyelle. Outre cette contextualisation, il est également possible d'attacher à chaque variante de prononciation une probabilité de telle sorte que la somme de toutes les variantes d'un même mot vale 1. Lorsque ramenée à l'échelle d'une séquence de mot, cette probabilité de prononciation s'intègre alors comme un nouveau terme dans l'équation (1.4). Enfin, pour des raison d'efficacité algorithmique, un lexique phonétisé est communément compilé sous la forme d'un *arbre lexical* dont un exemple est donné par la figure 1.2. À l'exception de la racine, chaque nœud de cet arbre représente un phonème et les arcs entre ces nœuds correspondent aux transitions entre phonèmes répertoriées dans le lexique phonétisé. En outre, les feuilles et certains autres nœuds de cet arbre correspondent à des fins de prononciations de certains mots. À chacun de ces nœuds est associée la liste des mots qui reconnaissent la séquence de phonèmes parcourue à partir de la racine comme une de leurs transcriptions phonétiques. Cette représentation est notamment utile pour le décodage acoustique du signal.

1.1.3 Modèle acoustique

D'après l'équation (1.3), le rôle du modèle acoustique est de calculer la vraisemblance $p(Y|W)$ du signal représenté par la séquence d'observations Y sachant une séquence de mots W. En pratique, ce calcul s'effectue *via* plusieurs décompositions s'appuyant sur un même formalisme, celui des modèles de Markov cachés. Cette décomposition est illustrée sur un exemple en figure 1.3.

À un premier niveau, la séquence de mots W se décompose en plusieurs séquences de phonèmes, concaténations des prononciations possibles de chacun de ses mots. Ces séquences sont représentées sous la forme d'un modèle de Markov caché où chaque état désigne la réalisation éventuelle d'un phonème et où les probabilités d'émission sont des fonctions de vraisemblance d'un phonème sachant le signal observé à un instant donné.

À un niveau plus fin, chaque phonème est généralement modélisé par un modèle de Markov caché à trois états qui modélisent respectivement le début, le milieu et la fin du

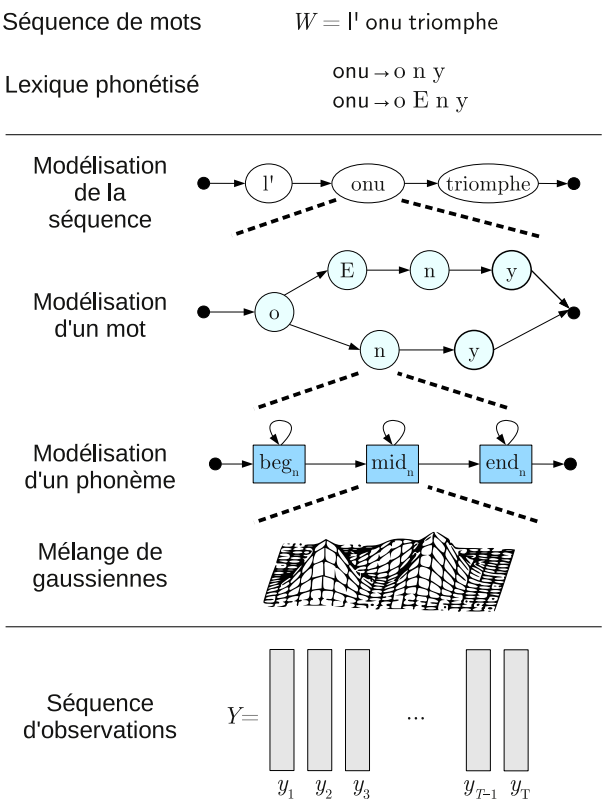


FIGURE 1.3 – Représentation d'une séquence de mots W sous la forme de modèles de Markov cachés pour le calcul de la vraisemblance $p(Y|W)$ du signal de parole Y.

phonème. Les probabilités d'émission de chacun de ces états sont généralement données par des mélanges de gaussiennes appartenant au même espace que les vecteurs acoustiques de Y.

La topologie de ces modèles étant fixée à l'avance, l'ensemble de ces probabilités d'émission et de transition ainsi que l'ensemble des gaussiennes peuvent être estimés à partir d'un corpus oral aligné avec sa transcription phonétique. Pour plus de détails sur la modélisation acoustique, le lecteur est invité à se référer à (Rabiner, 1989).

1.1.4 Modèle de langue

Le modèle de langue permet de calculer la probabilité *a priori* de séquences de mots. Dans le cadre qui nous intéresse des modèles de langue statistiques, cette probabilité est décomposée en probabilités conditionnelles, où la probabilité de chaque mot w_i de W est calculée sachant l'historique des mots $w_1 ... w_{i-1}$ supposés le précéder. Ceci s'exprime comme :

$$P[w_1, w_2, ...w_N] = P[w_1] \times \prod_{i=2}^{N} P[w_i|w_1...w_{i-1}] \ . \tag{1.5}$$

Typiquement, chacune de ces probabilités conditionnelles est estimée à partir d'un vaste corpus textuel d'apprentissage. Cependant, de manière évidente, cette estimation devient vite impossible ou beaucoup trop peu fiable lorsque le nombre de mots dans l'historique commence à être trop grand. Pour pallier ce problème, la stratégie la plus répandue consiste

à poser comme hypothèse que la prédiction d'un mot se fait indépendemment des mots qui lui sont les plus en amont dans la séquence considérée et à ainsi tronquer à gauche tout historique de manière à n'en conserver que les quelques mots les plus récents. La probabilité d'un mot w_i sachant ceux qui le précèdent s'approxime alors de la manière suivante :

$$P[w_i|w_1, ...w_{i-1}] \approx P[w_i|w_{\mathbf{i-n+1}}...w_{i-1})] \ , \tag{1.6}$$

où n est noté comme étant l'*ordre* du modèle et vaut typiquement entre 2 et 5. Les séquences de n mots $w_{i-n+1}...w_i$ sont appelées n-grammes et, par extension, on parle de probabilités n-grammes et de modèles de langue n-grammes. Plus communément, pour les cas où n vaut 1, 2, 3 ou 4, on emploie respectivement les termes *unigramme*, *bigramme*, *trigramme* et *quadrigramme*. Bien que, comme nous le verrons au chapitre 2, d'autres modélisations linguistiques soient possibles, les modèles n-grammes sont les plus répandus au sein des systèmes de reconnaissance automatique de la parole actuels.

Maintenant que nous avons posé les bases théoriques de la reconnaissance automatique de la parole, nous pouvons aborder la manière dont ces principes sont concrêtement mis en place au sein d'un système.

1.2 Décodage et questions algorithmiques

Partant de la maximisation décrite par l'équation (1.4) et des modèles à notre disposition, la question du décodage d'un signal de parole consiste à parcourir l'espace de recherche que représente l'intégralité des séquences de mots possibles à partir du vocabulaire du système et à calculer le score de chacune d'entre elles. Clairement, cet espace de recherche est infini puisqu'aucune hypothèse n'est faite sur la longueur des séquences à trouver. Quand bien même une telle hypothèse serait faite, le test de toutes les hypothèses de transcription possibles reste irréaliste étant donnée l'explosion combinatoire de leur nombre face à la taille du vocabulaire. Un système de reconnaissance s'appuie alors sur plusieurs stratégies qui, conjointement utilisées, permettent de maintenir des temps de décodage raisonnables.

1.2.1 Partitionnement du signal de parole

La durée du signal peut être un problème car, lorsqu'elle augmente, la longueur de la transcription attendue en sortie augmente également. Ainsi, des signaux d'entrée trop longs impliquent le parcours d'espace de recherche gigantesques et nécessitent des temps de traitements prohibitifs. En pratique, pour contourner cette limite, un signal de parole en entrée est découpé en segments relativement courts, typiquement d'une durée de 10 à 30 secondes. Pour ce faire, une frontière peut notamment être posée dans le signal à chaque fois qu'un locuteur marque une pause silencieuse de quelques dixièmes de seconde, par exemple pour prendre sa respiration. Par la suite, nous ferons référence à ces segments *via* le terme *groupes de souffle*. Une fois le signal segmenté, la transcription de celui-ci se réduit à transcrire successivement chacun de ses groupes de souffle.

1.2.2 Processus de transcription multi-passes

Les systèmes de reconnaissance automatique de la parole à grand vocabulaire fonctionnent souvent selon un processus itératif visant à réduire progressivement l'espace de recherche. Ce processus s'organise en plusieurs passes où, à l'issue de chacune, l'espace de

recherche courant est conservé et est transmis à la passe suivante. Chaque nouvelle passe consiste alors à re-parcourir l'espace de recherche courant en utilisant des modèles plus précis que ceux utilisés jusqu'alors, permettant ainsi d'écarter des hypothèses de transcription jugées trop peu prometteuses par les nouveaux modèles.

En pratique, les différents espaces de recherche manipulés sont représentés sous la forme de graphes dirigés acycliques, dits *graphes de mots*, dont les nœuds sont des instants du signal et dont chaque arc représente une hypothèse de mot à laquelle sont associés deux scores : un score acoustique correspondant à la vraisemblance du mot étant donnée la position courante dans le signal, score donné par le modèle acoustique ; et un score linguistique qui exprime la probabilité du mot sachant les mots qui le précèdent, score donné par le modèle de langue.

Le processus de reconnaissance d'un groupe de souffle comporte généralement au moins trois étapes : une première qui, sur la base d'un modèle de langue peu précis et d'un modèle acoustique générique, crée un premier graphe de mots pour chaque groupe de souffle ; une deuxième qui utilise les meilleures hypothèses de transcription de ces graphes de mots pour adapter le modèle acoustique à la voix du locuteur ; et enfin, une troisième étape qui se base sur un modèle de langue plus précis et des modèles acoustiques adaptés à la voix du locuteur. À l'issue de ces étapes, la taille des graphes de mots est telle qu'il devient envisageable de leur appliquer un algorithme de recherche du meilleur chemin pour obtenir la séquence de mots la plus probable. En pratique, les algorithmes généralement utilisés pour cette recherche sont l'algorithme de Viterbi (Viterbi, 1967) ou l'algorithme A* (Hart *et al.*, 1968).

1.2.3 Décodage en faisceau

Parmi les multiples étapes du processus de reconnaissance, la première d'entre elles, celle qui génère les premiers graphes de mots, est fondamentale car elle conditionne le déroulement de celles qui lui succèdent. Par ailleurs, il s'agit d'une étape coûteuse car elle s'attaque à l'espace de recherche complet. Pour ne pas avoir à parcourir cet espace en entier, les graphes de mots sont construits selon un algorithme dit de décodage en faisceau (*beam search*) qui s'appuie sur des principes de programmation dynamique (Ney, 1991). Cet algorithme se montre particulièrement intéressant car il propose une stratégie conjointe de parcours en largeur de l'espace de recherche et de calcul des scores des hypothèses de transcription. Par ailleurs, cet algorithme permet d'élaguer les hypothèses partielles les moins prometteuses au fur et à mesure du décodage, ce qui le rend très performant en terme de temps de calcul.

1.2.3.1 Parcours de l'espace de recherche et calcul des scores

L'algorithme de décodage en faisceau permet de construire progressivement un graphe de mots de manière trame-synchone, c'est-à-dire simultanément avec le parcours du signal. Comme présenté en figure 1.4, l'algorithme construit à la volée et met à jour en permanence un graphe d'états représentant l'ensemble des hypothèses de transcription. La construction de ce graphe s'effectue en parcourant des copies de l'arbre lexical, chacune de ces copies correspondant à une hypothèse partielle $w_1...w_i$ ou, plus exactement dans le cas d'un modèle n-gramme, à un historique $h = w_{i-n+1}...w_i$. Au sein de ces arbres, une nouvelle hypothèse partielle $w_1...w_{i+1}$ est définie à chaque fois qu'est atteint un état signifiant la fin de prononciation d'un mot w_{i+1}. Une nouvelle copie d'arbre correspondant

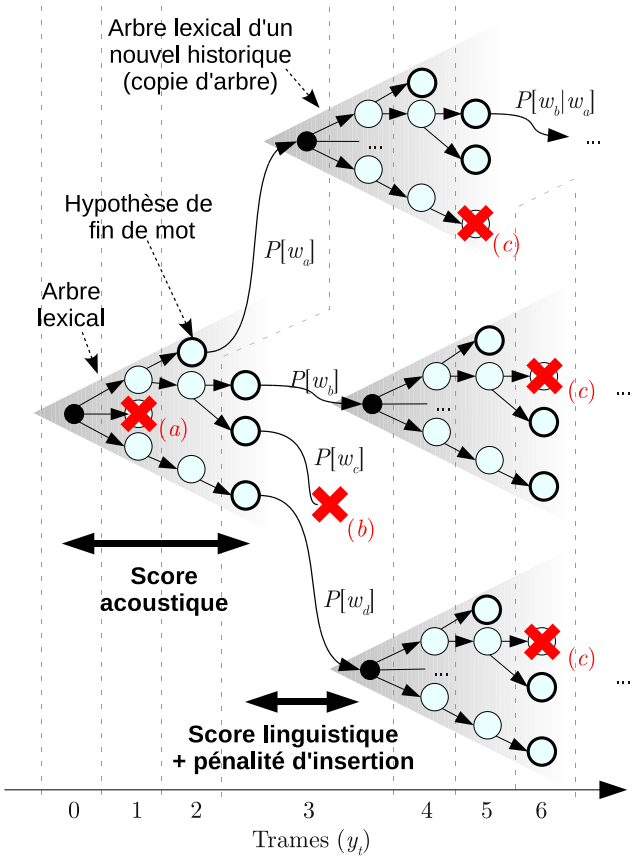


FIGURE 1.4 – Schéma du déroulement de l'algorithme de décodage en faisceau.

au nouvel historique $h' = w_{i-n+2}...w_{i+1}$ est alors créée afin de permettre la poursuite du décodage. Parallèlement à cette construction de l'espace de recherche, l'algorithme calcule, à chaque nouvelle trame y_t, un score $Q_h(s, y_t)$ pour chaque état s actuellement parcouru pour chaque historique actif h. Ce score est évalué à partir des scores des prédécesseurs de s et reflète ainsi l'expression partielle des produits des scores acoustiques et linguistiques calculés jusque là. Plus précisément, le score attribué à un nouvel état s est le maximum des scores obtenus à partir de chacun de ses prédécesseurs. Dans cette logique, deux cas se distinguent. Si s est un état quelconque de l'arbre lexical d'un historique h, le calcul du nouveau score se fait uniquement grâce au modèle acoustique :

$$Q_h(s, y_t) = \max_{s' \to s} \left\{ Q_h(s', y_{t-1}) \times p(s|y_t) \times P(s' \to s) \right\} \ , \tag{1.7}$$

où $s' \to s$ signifie que s' est un prédecesseur de s, $p(s|y_t)$ est la vraisemblance de l'état s sachant la trame y_t et $P(s' \to s)$ est la probabilité de transition entre s' et s. Si s correspond à l'hypothèse de fin de prononciation d'un mot w_{i+1}, il faut en outre associer un score à la racine s_0 du nouvel arbre lexical créé. Ce score reprend le nouveau score calculé pour s et intègre la probabilité $P[w_{i+1}|h]$ fournie par le modèle de langue, ainsi que la pénalité I d'insertion d'un nouveau mot dans l'hypothèse partielle courante :

$$Q_{h'}(s_0, y_t) = \max_{h \to h'} \left\{ Q_h(s, y_t) \times P[w_{i+1}|h]^{\psi} \times I \right\} \ , \tag{1.8}$$

où $h \rightarrow h'$ signifie que l'historique h conduit, après l'hypothèse de fin de mot de w_{i+1}, à l'historique de même ordre $h' = w_{\mathbf{i-n+2}}...w_{\mathbf{i+1}}$ correspondant au nouvel arbre lexical et ψ est le facteur de mise à l'échelle des probabilités du modèle de langue. Finalement, l'algorithme de décodage se termine quand il n'y a plus aucun nouvel état à parcourir, par exemple dans le cas où toutes les trames ont été « consommées » par l'algorithme.

1.2.3.2 Stratégies d'élagage

Dans sa version brute, telle que nous venons de la présenter, l'algorithme de décodage en faisceau demeure d'une complexité calculatoire élevée puisque toutes les possibilités de transcriptions sont évaluées. Heureusement, il est possible d'y intégrer une stratégie de suppression des hypothèses partielles les moins prometteuses en élaguant le graphe d'états en même temps que celui-ci est construit. Cet élagage s'effectue par des contraintes sur les scores de chaque état. Trois types principaux de contraintes sont considérés. Des exemples de l'effet des contraintes de chacun de ces types sont présentés sur la figure 1.4 par les légendes (a), (b) et (c). Premièrement, à chaque trame y_t du signal décodé, seules sont conservés les états dont le score est suffisamment proche du meilleur score courant parmi l'ensemble des copies d'arbre (légende a). Pour chaque nouvel état s envisageable pour l'arbre lexical de l'historique h, cette contrainte d'existence s'exprime comme :

$$Q_h(s, y_t) > \delta_{AC} \times \max_{(h',s')} \left\{ Q_{h'}(s', y_t) \right\} \tag{1.9}$$

où la constante $\delta_{AC} \in [0, 1]$, appelé *seuil acoustique*, définit la tolérance acoustique du système. Deuxièmement, le même type de seuillage est appliqué pour chaque hypothèse de fin de mot, c'est-à-dire que seules seront explorées les nouvelles copies d'arbre des historiques dont le score est suffisamment proche du score $Q_{ML}(t)$ de la meilleure hypothèse de fin de mot à la trame y_t (légende b). La contrainte est alors :

$$Q_{h'}(s_0, y_t) > \delta_{ML} \times Q_{ML}(t) \ , \tag{1.10}$$

où le facteur $\delta_{ML} \in [0, 1]$ est appelé *seuil linguistique*. Enfin, une dernière contrainte restreint l'ensemble des hypothèses actives à chaque trame aux seules M hypothèses ayant le meilleur score (légende c). En pratique, les constantes δ_{AC}, δ_{ML} et M sont des valeurs empiriques obtenues par la recherche d'un compromis optimal entre différents critères comme la qualité des transcriptions automatiques générées, la durée de cette génération et la taille des graphes de mots. Cet apprentissage est important car il définit le niveau de tolérance du système de reconnaissance au niveau acoustique et linguistique, tolérance ayant une influence directe sur la qualité des sorties du système.

Maintenant que nous avons dévoilé les aspects théoriques et pratiques du fonctionnement d'un système de reconnaissance automatique de la parole, voyons les différents types de sortie qu'il peut produire.

1.3 Sorties d'un système de reconnaissance automatique de la parole

Un système de reconnaissance automatique de la parole peut fournir différentes sorties. Outre la transcription finale ou les quelques meilleures hypothèses de transcription trouvées

par le système, il peut être intéressant de travailler sur des sorties intermédiaires, telles que les graphes de mots, ou de tirer parti d'informations complémentaires calculées *a posteriori*, comme des mesures de confiance. Cette section présente ces différentes sorties.

1.3.1 Hypothèses de transcription

La sortie par défaut d'un système de reconnaissance est la transcription textuelle du signal de parole d'entrée, c'est-à-dire la séquence de mots considérée comme la plus probable par le système de transcription. En tant que telle, cette sortie peut être considérée comme du texte et il devient alors envisageable d'y appliquer des techniques venant du monde du traitement des textes écrits. Toutefois, certaines caractéristiques distinguent les transcriptions d'un texte classique. Tout d'abord, les transcriptions automatiques ne contiennent aucune ponctuation. Ceci s'explique principalement par le fait que la notion de ponctuation est difficilement transposable à la parole et que, par conséquent, les signes de ponctuation sont difficilement modélisables et intégrables au sein du processus de décodage. Le découpage d'une transcription en propositions voire en phrases est ainsi peu envisageable. Ensuite, une majorité de systèmes de reconnaissance automatique de la parole ne modélise pas la casse des mots, conduisant ainsi à des transcriptions entièrement en minuscule. Cette particularité s'explique par le fait que la prise en compte de la casse augmente considérablement la taille du vocabulaire, et donc les temps de décodage, mais ajoute également de l'ambiguïté dans la modélisation linguistique. Cette absence de casse est notamment préjudiciable pour des tâches d'extraction d'entités nommées[2]. Enfin, les transcriptions présentent surtout l'inconvénient majeur d'inclure potentiellement des mots mal transcrits. Selon la difficulté de la tâche de transcription, la proportion de ces mots peut grandement varier. Par exemple, avec de bonnes conditions acoustiques et dans le cadre d'un discours dont l'énoncé est bien modélisé par le modèle de langue, certaines transcriptions peuvent ne contenir qu'un nombre marginal de fautes. À l'inverse, lorsque les conditions acoustiques sont difficiles et que l'énoncé du discours diffère grandement de ce pour quoi le modèle de langue a été appris, on peut parfois frôler 100 % de mots mal transcrits. C'est par exemple le cas lorsqu'on applique à des retransmissions sportives un système de reconnaissance automatique de la parole appris pour transcrire des journaux d'actualités. Dans ces conditions, il devient alors difficile, voire impossible, d'envisager un quelconque traitement à partir des transcriptions fournies par le système.

Une alternative à la manipulation d'une simple transcription, unique meilleure séquence trouvée par le système, peut alors être de se baser sur la liste ordonnée des quelques $\mathcal{N}$ meilleures séquences issues du décodage d'un graphe de mots. Ces listes présentent comme intérêt d'offrir plus de richesse que les transcriptions et de pouvoir être réordonnées en prenant en compte des sources d'information que le système de transcription ne considérait jusqu'alors pas, comme, par exemple, la syntaxe (Kaufmann *et al.*, 2009) ou la morphosyntaxe[3] (Huet *et al.*, 2010).

[2] C'est-à-dire l'extraction des noms de personnes, de lieux, d'organisations...

[3] La morphosyntaxe s'intéresse aux catégories grammaticales (nom, verbe, adjectif...) et à l'information flexionnelle (genre, nombre, personne...) des mots dans une phrase.

1.3.2 Graphes de mots et réseaux de confusion

Les graphes de mots représentent sous une forme plus ou moins compacte l'espace de recherche à un instant donné du processus multi-passes de décodage. Pour un groupe de souffle donné, cette structure consiste en un graphe dont les nœuds sont des instants du signal et où les arcs représentent des hypothèses de mots accompagnés de leur vraisemblance acoustique et de leur probabilité linguistique. Les graphes de mots sont donc des objets intéressants car ils contiennent beaucoup d'informations. Cependant, ils peuvent être très gros en terme de nombre de nœuds et d'arcs et devenir difficilement manipulables. Il devient alors intéressant de les élaguer ou de les compacter, par exemple sous la forme de réseaux de confusion.

Les réseaux de confusion peuvent être vus comme des graphes de mots dont certains nœuds ont été fusionnés en alignant temporellement les meilleures hypothèses issues d'un graphe de mots. Il en résulte un graphe linéaire dont les nœuds délimitent des tranches de temps du signal. Les arcs entre deux nœuds successifs quelconques représentent alors les hypothèses de mots en concurrence pour chaque tranche de temps. Ces arcs sont pondérés par la probabilité *a posteriori* de chaque mot w, cette dernière étant fonction de la somme des probabilités de tous les arcs du graphe de mots d'origine qui appartenait à la tranche de temps considérée et qui portait le mot w. Les réseaux de confusion sont des objets intéressants car ils permettent d'introduire des hypothèses de transcription qui n'existaient pas dans les graphes de mots. Grâce à une technique dite *décodage par consensus*, ces nouvelles hypothèses permettent notamment d'améliorer la qualité des transcriptions finales d'un système de reconnaissance automatique de la parole en sélectionnant, pour chaque intervalle de temps, le mot qui maximise sa probabilité *a posteriori* (Mangu *et al.*, 2000).

1.3.3 Mesures de confiance

En sus des hypothèses de transcription que fournit un système de reconnaissance automatique de la parole, il est possible de calculer, pour chaque mot d'une hypothèse, un *score de confiance* compris entre 0 et 1, ce score étant d'autant plus élevé que le système estime fiable la transcription du mot. Ces scores sont typiquement calculés à partir des probabilités *a posteriori* ou d'informations dérivées des graphes de mots (Wessel *et al.*, 2001). Ils peuvent aussi être issus de processus plus sophistiqués faisant, par exemple, appel à des techniques d'apprentissage artificiel prenant en compte des informations autres que les seuls scores acoustiques et linguistiques fournis par le système (Fayolle *et al.*, 2010). Il est à noter toutefois que les mesures de confiance ne sont elles-mêmes par toujours fiables et qu'elles doivent donc être considérées avec prudence.

La prochaine section présente des méthodes d'évaluation des différents types de sorties que nous venons de présenter.

1.4 Méthodes d'évaluation

De par la multitude de composants impliqués dans le fonctionnement d'un système de reconnaissance automatique de la parole et la variété des types de sorties, il existe un grand nombre de méthodes pour évaluer la qualité de l'intégralité ou d'une partie d'un système de transcription. Dans cette section, nous ne présentons que les mesures qui nous intéressent

pour notre travail d'adaptation thématique et que nous serons amené à évoquer dans la suite de ce manuscrit.

1.4.1 Taux d'erreur d'une transcription

Étant donné que les systèmes de reconnaissance ont pour vocation de générer des transcriptions textuelles, les performances générales de ces systèmes se mesurent très généralement par des taux d'erreur relatifs au nombre de mots mal transcrits dans les séquences de mots retournées par rapport à une transcription de référence, le plus souvent d'origine manuelle. Plus ces taux sont bas, plus le système est considéré comme performant. Ces taux sont calculés en alignant les transcriptions automatiques du système avec leurs transcriptions de référence de manière à minimiser une distance d'édition entre ces séquences respectives (Levenshtein, 1966). Cette distance d'édition est fonction du nombre d'opérations élémentaires – substitution d'un mot par un autre, insertion d'un mot et suppression d'un mot – à effectuer pour passer de la séquence de référence à la séquence retournée par le système. Ce principe générique conduit à diverses mesures.

Taux d'erreur sur les mots. Le taux d'erreur sur les mots d'une transcription se calcule comme la distance minimale d'édition entre cette transcription et la transcription de référence, rapportée au nombre de mots de la référence. Cette mesure se note *WER*, pour l'anglais *word error rate*, et s'exprime comme le pourcentage suivant :

$$\mathrm{WER} = \frac{N_{\mathrm{sub}} + N_{\mathrm{ins}} + N_{\mathrm{sup}}}{N_{\mathrm{total}}} \ , \tag{1.11}$$

où N_{sub}, N_{ins} et N_{sup} sont respectivement le nombre d'opérations de substitution, d'insertion et de suppression impliquées dans le calcul de la distance minimale d'édition, et N_{total} est le nombre de mots dans la transcription de référence. Il s'agit de la mesure la plus communément utilisée car elle donne une idée générale des performances du système. En alternative au WER, il n'est cependant pas rare de voir les performances d'un système exprimées *via* un taux dit de reconnaissance s'exprimant comme $1 - WER$. D'autres taux d'erreur existent, souvent dédiés à l'évaluation d'un système pour des tâches spécifiques.

Variantes du taux d'erreur sur les mots. Il est possible de restreindre le WER à un sous-ensemble de mots porteurs d'intérêt pour une tâche donnée plutôt que de s'intéresser à l'ensemble des mots. Un tel taux d'erreur peut se calculer soit à partir des transcriptions automatiques et de référence desquels ont préalablement été supprimés tous les mots jugés sans importance, soit en pondérant les mots de la référence en fonction de leur importance pour la tâche considérée de manière à ne pas pénaliser de la même façon toutes les erreurs de transcription (Nanjo *et al.*, 2005). D'autres traitements peuvent être effectués préalablement au calcul du taux d'erreur. Il est notamment possible de lemmatiser les transcriptions, c'est-à-dire de ramener chacun de leurs mots vers une forme canonique, appelée *lemme*[4], afin de faire abstraction des erreurs de flexion pouvant être faites par le système de reconnaissance, notamment entre flexions homophones. Dans notre travaux, nous seront amenés

[4]Par exemple, le lemme d'un verbe conjuguée peut être sa forme infinitive, celui d'un adjectif féminin pluriel sa forme masculin singulier...

à combiner ces deux variantes au sein du calcul d'un taux d'erreur sur les lemmes des mots lexicaux[5] que nous notons LER pour *lemma error rate*.

Taux d'erreur sur les phrases. Les performances globales d'un système de reconnaissance peuvent se rapporter à un grain plus gros que celui du mot par le calcul du pourcentage de groupes de souffle mal transcrits par rapport à la référence, c'est-à-dire le pourcentage de groupes de souffle de la transcription automatique qui contiennent au moins une erreur par rapport au nombre total de groupes de souffle dans la référence. Ce taux d'erreur est noté *SER*, pour *sentence error rate*, littéralement « taux d'erreur sur les phrases » bien que, comme nous l'avons déjà évoqué, la notion de phrase n'existe pas vraiment pour le langage transcrit. Ce taux d'erreur s'avère intéressant pour des applications où tous les mots d'une transcription sont importants. Par exemple, dans le cadre d'un dialogue homme-machine (système de questions/réponses, système de réservation...), il est important que le système de reconnaissance transcrive bien les marques de négation là où il y en a et n'en ajoute pas là où il n'y en a pas.

1.4.2 Taux d'erreur des graphes de mots : un taux oracle

Un taux d'erreur oracle, ou simplement taux oracle, désigne le plus petit taux d'erreur sur les mots que l'on peut atteindre en théorie à un moment donné du processus de transcription. De manière relativement répandue, un tel taux est calculé à partir des graphes de mots construits par le système pour un document à transcrire – on parle alors *graph error rate*, ou *GER* – en alignant chaque graphe avec la transcription de référence du groupe de souffle qui lui correspond. Cet alignement s'effectue en cherchant le chemin du graphe de mots qui minimise la même distance d'édition que décrite précédemment. Alors que les taux d'erreur sur les transcriptions n'évaluent que les performances finales d'un système de reconnaissance automatique de la parole, le taux oracle sur les graphes de mots a l'intérêt de donner une idée du potentiel de reconnaissance à un instant donné du processus de transcription. Il est toutefois à interpréter avec prudence puisque plus un graphe de mots est gros, plus il a de chances de recenser la bonne hypothèse de transcription[6]. Ainsi, pour être comparables, des taux oracle doivent être calculés sur des graphes de mots de même taille.

1.4.3 Qualité des mesures de confiance

Comme évoqué en 1.3.3, les mesures de confiance renvoyées par un système ne sont pas parfaitement fiables. Différentes mesures existent pour évaluer ce manque de fiabilité. Dans l'ensemble, le calcul de ces indices de fiabilité se fonde sur celui du pouvoir prédictif des mesures de confiance pour distinguer les mots mal transcrits de ceux correctement transcrits au sein d'une transcription automatique. Ce pouvoir prédictif varie d'un indice de fiabilité à l'autre ; il peut se présenter sous la forme d'une erreur de classification (*confidence error rate* ou *CER*) ou se fonder sur des mesures issues de la théorie de l'information (*normalized*

[5] Les mots lexicaux, ou mots pleins, représentent l'ensemble des mots de la langue à l'exception de ceux à l'usage le plus courant (préposition, articles, auxilaires...). Ces derniers, appelés *mots vides*, sont déconsidérés car ils sont jugés non significatifs pour la description du contenu d'un texte.

[6] La présence de la bonne hypothèse dans le graphe de mots ne garantit néanmoins en rien que celle-ci soit jugée par le système comme la plus probable. Ainsi, augmenter la taille des graphes de mots ne présente pas d'intérêt en soi pour améliorer la qualité des transcriptions finales du système, d'autant plus que cette augmentation se traduit par des temps de calcul accrus.

cross entropy ou NCE).

Après avoir donné une vue d'ensemble du fonctionnement d'un système reconnaissance automatique de la parole à grand vocabulaire, nous terminons cette section par la description plus particulière du système de transcription IRENE, système utilisé dans tous nos travaux.

1.5 Le système IRENE

Le système IRENE est un système de reconnaissance automatique de la parole développé pour le français à l'IRISA sur la base de travail initié conjointement avec Telecom Paris. Il s'agit du système actuellement utilisé à l'IRISA et sur lequel s'est appuyée l'intégralité des expériences relatées dans ce manuscrit.

Le système IRENE s'appuie sur les composants suivants :

- un vocabulaire de 65 000 mots pour un total de 80 000 prononciations dans le lexique phonétisé ;
- 2 modèles de langue n-grammes fondés sur ce vocabulaire : un modèle trigramme et un modèle quadrigramme, plus précis ;
- 2 types de modèles acoustiques génériques : l'un dit monophone, c'est-à-dire dont l'estimation de la vraisemblance d'un phonème ne dépend pas des phonèmes qui l'entourent, l'autre triphone, c'est-à-dire, à l'inverse, qui tire parti du contexte phonémique. Ce dernier modèle ne gère toutefois pas les liaisons entre mots. Par défaut, il existe deux déclinaisons pour chaque modèle acoustique : l'une pour les hommes, l'autre pour les femmes.

Ces différents composants ont été appris à partir des données d'apprentissage de la première campagne d'évaluation ESTER (Galliano *et al.*, 2005). Ces données rassemblent des transcriptions manuelles de 70 heures de journaux d'actualités diffusés entre 1998 et 2003 sur différentes radios francophones. En sus de ces données, l'apprentissage du vocabulaire et des modèles de langue s'est également appuyé sur des articles d'archives du journal *Le Monde* entre 1987 et 2002. Plus de détails sur l'apprentissage des composants linguistiques du système sont donnés au chapitre 2.

Fondé sur ces composants, le processus de transcription du système IRENE s'effectue grâce au décodeur SIROCCO (Gravier *et al.*, 2002). À partir d'un découpage préalable du signal par le système en locuteurs et en groupes de souffle, le décodage se déroule en quatre étapes schématisées par la figure 1.5. Une première étape consiste à créer des graphes de mots pour chaque groupe de souffle sur la base des modèles acoustiques monophones et le modèle de langue trigramme. Dans un second temps, ces graphes de mots sont étendus et réévalués en utilisant des modèles acoustiques triphones et le modèle de langue quadrigramme. Lors d'une troisième étape, les modèles acoustiques triphones sont adaptés pour chaque locuteur en s'appuyant sur une transcription obtenue par l'algorithme de Viterbi. Ces modèles plus précis permettent d'élaguer les graphes de mots après leur ré-évaluation. Dans un dernier temps, la liste des $\mathcal{N}$ meilleures hypothèses est générée à partir des graphes de mots élagués. Ces hypothèses sont ré-ordonnées de manière à favoriser les séquences de mots les plus cohérentes sur le plan morphosyntaxique, c'est-à-dire celles dont les parties du discours s'enchaînent le plus probablement d'après un modèle n-gramme (Huet *et al.*, 2007). À l'issue de cette étape, un décodage par consensus permet d'obtenir la transcription finale.

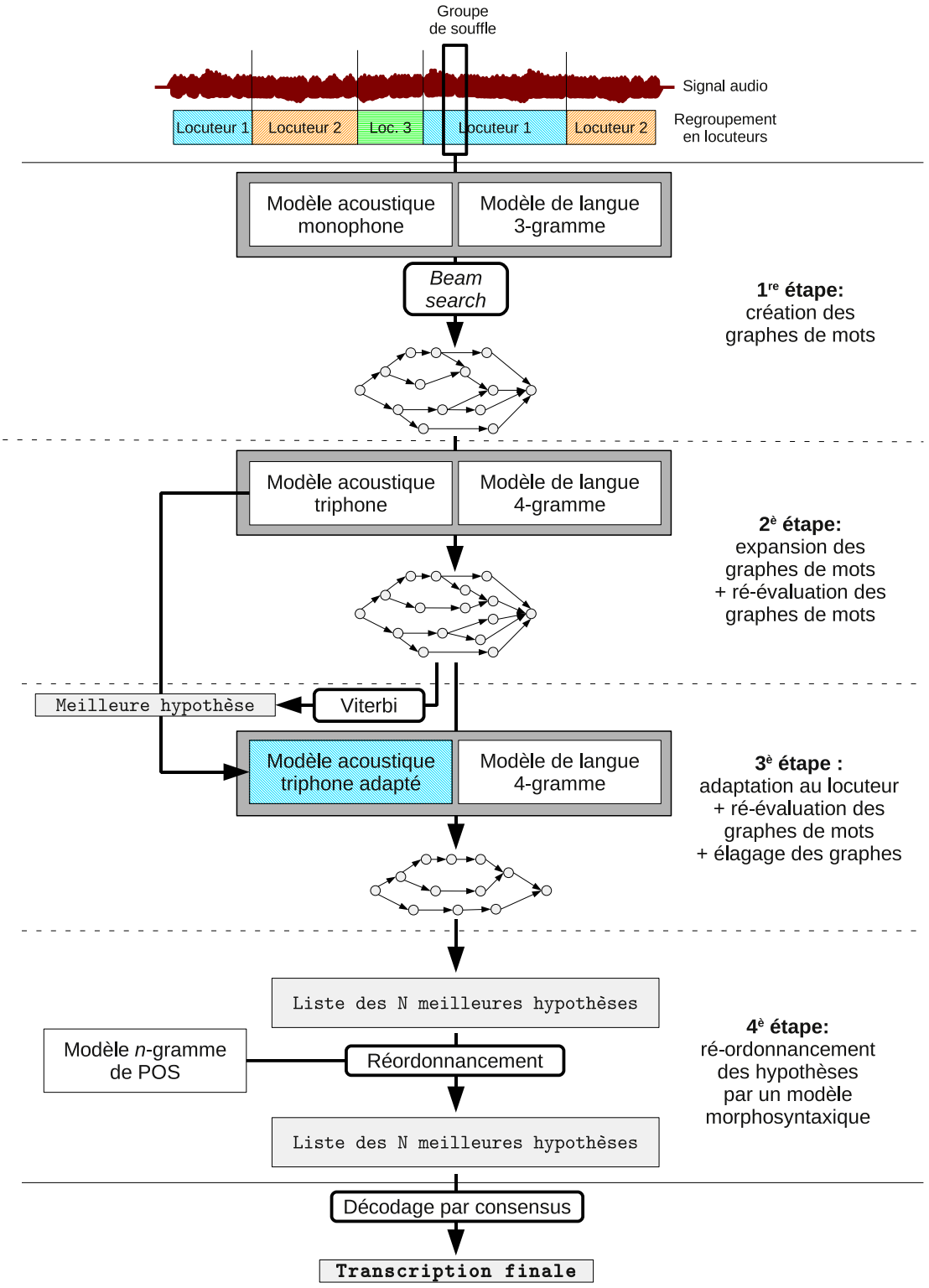


FIGURE 1.5 – Processus multi-passes de transcription d'un groupe de souffle au sein du système IRENE.

Les performances du système IRENE ont été évaluées dans le cadre de la campagne ESTER. Pour un temps de décodage d'environ 8 fois le temps réel – il faut donc 8 heures pour transcrire 1 heure de parole – , le système atteint un taux d'erreur sur les mots d'environ 20 % pour un taux oracle des graphes de mots à la fin de la première étape de 9-10 %. Ces performances, dans la moyenne des systèmes de transcription actuellement proposés pour le français, permettent d'envisager l'utilisation dans un cadre expérimental des sorties fournies par le système de l'IRISA.

Ce chapitre nous a permis de présenter le fonctionnement général d'un système de reconnaissance automatique de la parole continue à grand vocabulaire, en détaillant en particulier le processus mis en œuvre par le système IRENE utilisé dans nos travaux. Au sein de ce processus, le prochain chapitre s'attarde sur le point particulier qui nous intéresse, à savoir la modélisation linguistique du langage. Nous souhaitons particulièrement souligner en quoi la traditionnelle modélisation fondée sur un vocabulaire fermé et sur un modèle de langue n-gramme souffre de lacunes pour transcrire correctement des documents tout-venants traitant de thèmes variés et que le recours à des techniques d'adaptation thématique de ces composants est donc nécessaire.

Chapitre 2

Modélisation statistique du langage

La modélisation statistique d'un langage s'appuie sur deux éléments essentiels : un vocabulaire, qui décrit l'ensemble des unités lexicales de la langue considérée, et un modèle de langue dont le rôle est de représenter les possibilités d'enchaînement des mots du vocabulaire. Il s'agit d'une problématique partagée par de nombreux domaines. Outre le cas de la reconnaissance automatique de la parole (*cf.* chapitre 1), le recours à une modélisation statistique du langage intervient, par exemple, en traduction automatique, en recherche d'information, en catégorisation de texte ou encore en reconnaissance optique de caractères.

Du côté du vocabulaire, le choix des unités lexicales sur lesquelles repose un modèle de langue reflète différents enjeux. D'une part, il doit permettre de couvrir au mieux l'usage de la langue pour les tâches pour lesquelles le langage doit être modélisé. À titre d'exemple, en reconnaissance automatique de la parole, si cette couverture est trop faible, de nombreux mots seront vraisemblablement mal transcrits, faute d'être modélisés dans le système. D'autre part, la nature des unités lexicales doit permettre au modèle de langue de modéliser efficacement des phénomènes linguistiques complexes tels que la composition.

Les modèles de langue statistiques visent à assigner des probabilités à des séquences de mots du vocabulaire par l'intermédiaire d'estimations faites sur un très vaste corpus textuel d'apprentissage. Ces approches permettent alors d'ordonner des hypothèses de transcription proches sur le plan acoustique. Cette approche se distingue des approches structurelles qui visent à modéliser le langage par un ensemble de règles d'une grammaire et qui ne permettent ainsi que de diagnostiquer si, oui ou non, une séquence appartient au langage modélisé. Parmi les nombreuses approches statistiques de la littérature, celles fondées sur l'estimation de probabilités de n-grammes, séquences de n mots, sont reconnues comme étant les plus performantes en reconnaissance automatique de la parole depuis leur introduction par (Jelinek, 1976). Les performances des modèles n-grammes surprennent pourtant de par la simplicité avec laquelle le langage naturel y est modélisé. Les notions de structure et de sémantique de la langue n'y sont notamment pas explicitement présentes. Pour combler ces absences, d'autres types de modèles de langues ont été proposés dans la littérature. Toutefois, les taux de reconnaissance engendrés par ces techniques plus sophistiquées au mieux égalent ceux du modèle n-gramme, pour un coût d'apprentissage paradoxalement généralement plus élevé.

L'objectif principal de ce chapitre est de montrer en quoi la modélisation linguistique fondée sur un vocabulaire fini et sur un modèle de langue n-gramme – modélisation à

laquelle nous nous intéressons car elle est la plus largement répandue dans les systèmes de transcriptions actuels, dont le nôtre – souffre d'un manque d'adaptabilité pour transcrire efficacement des documents traitant chacun d'un thème différent. Pour comprendre ce problème, nous détaillons le fonctionnement des modèles n-grammes. Nous exposons alors différentes tentatives de modélisation qui s'appuient sur les principes de fonctionnement du modèle n-gramme mais qui, en outre, cherchent à pallier le manque de connaissances linguistiques de ce dernier. Bien que globalement moins performantes que les modèles n-grammes pour la reconnaissance automatique de la parole, ces techniques introduisent des principes intéressants qui sont à l'origine de différentes techniques d'adaptation que nous présentons au chapitre 3 et qui nous permettront également d'ouvrir des perspectives de travail au cours de l'exposé de nos contributions. Suite à ces présentations, nous exposons en section 2.2 la méthode d'évaluation classique d'un modèle de langue statistique et dévoilons en section 2.3 le processus d'apprentissage d'un vocabulaire et d'un modèle de langue n-gramme. Finalement, à la section 2.4, nous soulignons les faiblesses de ces composants, notamment pour modéliser les variations linguistiques liées à des changements de thèmes. Ces faiblesses mettent en lumière les objectifs à atteindre pour le développement d'un processus adaptation thématique d'un vocabulaire et d'un modèle de langue n-gramme.

2.1 Modèles de langue existants

Pour rappel, le but d'un modèle de langue est de calculer la probabilité *a priori* d'une séquence de mots $W = w_1...w_N$ par la décomposition suivante :

$$P[w_1, w_2, ...w_N] = P[w_1] \times \prod_{i=2}^{N} P[w_i|w_1...w_{i-1}] \ , \tag{2.1}$$

où la séquence $w_1...w_{i-1}$ est appelé l'*historique*. Par défaut, ces probabilités conditionnelles sont calculées empiriquement par une technique de maximum de vraisemblance telle que

$$P[w_i|w_1...w_{i-1}] = \frac{C(w_1...w_i)}{\sum_v C(w_1...w_{i-1}v)} \ , \tag{2.2}$$

où $C(w_1...w_i)$ est le nombre d'occurrences de la séquence $w_1...w_i$ dans un vaste corpus d'apprentissage.

Selon ce principe, la probabilité de longues séquences devient vite nulle car aucun corpus d'apprentissage n'est suffisamment grand pour permettre d'observer toutes les séquences de mots composables à partir du vocabulaire, quand bien même ce dernier est fini. Il a alors été proposé de regrouper les historiques des probabilités conditionnelles en classes d'équivalence (Jelinek, 1998), la nature de ces classes étant arbitraire et hautement variable d'un modélisation à l'autre. En notant Φ la fonction qui associe à un historique sa classe d'équivalence, la formule (2.1) s'approxime alors par :

$$P[w_1, w_2, ...w_N] \approx P[w_1] \times \prod_{i=2}^{N} P[w_i|\mathbf{\Phi}(w_1...w_{i-1})] \ , \tag{2.3}$$

Le choix de cette fonction définit en grande partie les propriétés d'une modélisation.

Nous détaillons dans un premier temps le fonctionnement d'un modèle de langue n-gramme, puis exposons plus brièvement différentes autres modélisations présentes dans la littérature.

2.1.1 Le modèle *n*-gramme

Les modèles de langue n-grammes sont historiquement les premiers modèles de langue statistiques à avoir été proposés (Jelinek, 1976). Pourtant, depuis 30 ans, ils demeurent les modèles les plus performants pour la tâche de reconnaissance automatique de la parole. Ces modèles n-grammes présentent l'intérêt de s'intégrer efficacement dans un algorithme de décodage en faisceau et de se fonder sur une modélisation simple du langage qui nécessite, en théorie, un apprentissage tout aussi simple. En pratique toutefois, différents aménagements sont nécessaires pour combler certaines faiblesses du processus d'estimation des probabilités n-grammes par maximum de vraisemblance. Cette section rappelle le principe général d'un modèle n-gramme avant d'aborder ces considérations supplémentaires.

2.1.1.1 Principe

Les modèles n-grammes reposent sur l'hypothèse markovienne consistant à dire que la probabilité d'apparition d'un mot dépend uniquement de l'historique des quelques $n-1$ mots qui le précédent, cet historique étant souvent noté h. Plus formellement, cela revient à écrire que

$$P[w_i|\Phi(w_1...w_{i-1})] = P[w_i|h] = P[w_i|w_{\mathbf{i\text{-}n+1}}...w_{i-1}] \ . \tag{2.4}$$

Dans le cadre d'un système de reconnaissance automatique de la parole à grand vocabulaire, n, l'ordre du modèle, est typiquement compris entre 2 et 5. Cette hypothèse markovienne présente l'intérêt de poser une borne maximale au nombre d'historiques possibles et donc au nombre de probabilités à estimer – cette borne étant égale à $|V|^n$ si $|V|$ est la taille du vocabulaire. Pour cette même raison, l'utilisation d'un modèle de langue n-gramme contribue également à rendre particulièrement rapide le décodage d'un signal de parole en limitant la liste des historiques actifs que l'algorithme *beam search* doit entretenir lors du parcours de l'espace de recherche de toutes les hypothèses de transcription possibles.

Au delà de ce principe général, l'estimation de certaines probabilités n-grammes peut être rendue trop approximative par la technique de maximum de vraisemblance, ce qui nécessite quelques aménagements.

2.1.1.2 Lissage des probabilités *n*-grammes

L'application de la formule (2.2) du maximum de vraisemblance conduit à attribuer une probabilité nulle à tout n-gramme n'ayant jamais été rencontré dans le corpus d'apprentissage, quand bien même ce n-gramme pourrait être parfaitement valide sur le plan linguistique. Plus généralement, l'estimation par maximum de vraisemblance tend à sous-estimer les probabilités des n-grammes les moins fréquents dans le corpus d'apprentissage.

Ce problème dit d'*éparsité des données* est fortement marqué dans le cadre de l'apprentissage d'un modèle de langue car il a été montré que les mots et séquences de mots d'un langage naturel suivent la loi de Zipf qui consiste à dire que la fréquence d'apparition d'un mot décroît à un rythme inversement proportionnel à son rang d'apparition (Manning et Schütze, 1999 ; Ha *et al.*, 2002). Ainsi, une majorité de mots apparaissent peu de fois dans le corpus d'apprentissage et de nombreuses probabilités n-grammes sont donc peu fiables si l'on s'en tient aux estimations par maximum de vraisemblance[7]. Dans le cadre d'un

[7]Notons que ce problème d'éparsité des données n'est pas lié à la modélisation n-gramme ; il s'agit d'un problème récurrent dès lors que l'on cherche à estimer des probabilités pour le langage naturel à partir de fréquences d'apparition.

système de reconnaissance automatique de la parole, ce phénomène tend à condamner ces n-grammes à ne jamais, ou quasi jamais, apparaître dans les transcriptions finales.

Pour remédier à ce problème, des techniques dites de *lissage* proposent de modifier la distribution empirique des probabilités n-grammes en biaisant les comptes des séquences observées dans le corpus d'apprentissage de manière à réduire les probabilités des n-grammes plus fréquents. La masse de probabilité ainsi récupérée est alors redistribuée aux évènements rares ou n'ayant jamais été observés. Les plus notables de ces techniques sont le lissage de Good-Turing (Good, 1953), de Witten-Bell (Witten et Bell, 1991) et de Kneser-Ney (Kneser et Ney, 1995). Pour une étude détaillée des techniques de lissage actuelles, nous recommandons la lecture de (Chen et Goodman, 1998).

2.1.1.3 Techniques de repli

Puisque les techniques de lissage permettent entre autres de réserver une masse de probabilité aux n-grammes non observés, il est légitime de s'interroger sur la manière de redistribuer cette masse. Naïvement, sous l'hypothèse d'une distribution uniforme des n-grammes non observés, une première approximation de la probabilité jointe d'un de ces n-grammes hw peut s'exprimer comme :

$$P[hw] = \frac{1 - \sum_{h'w' \in E} P^*[h'w']}{|V|^n - |E|} \ , \tag{2.5}$$

où E est l'ensemble des n-grammes observés, P^* est la probabilité jointe fournie par le modèle de langue après lissage, V est le vocabulaire et $|.|$ désigne la cardinalité. Cette approximation est toutefois grossière car elle ne tient compte en rien des mots qui composent hw.

Pour pallier ce problème, une stratégie dite de *repli* (*backoff* en anglais) propose de répartir les masses de probabilité des n-grammes de fréquence nulle non pas globalement mais historique par historique (Katz, 1987). La probabilité d'un n-gramme $hw = w_{i-n+1}...w_{i-1}w_i$ s'écrit alors dans le cas général sous la forme récursive suivante :

$$P[w|h] = \begin{cases} P^*[w|h] & \text{si } hw \in E \\ \beta(h) \times P[w|h^-] & \text{sinon} \ , \end{cases} \tag{2.6}$$

où h^- est l'historique d'ordre inférieur $w_{i-n+\mathbf{2}}...w_{i-1}$ et $\beta(h)$ est appelé *coefficient de repli*. Ce coefficient peut se voir comme une mise à l'échelle des probabilités de l'historique h pour l'historique h^-. Il se calcule comme le rapport entre la masse de probabilité non encore allouée pour l'historique h et celle laissée libre pour l'historique h^- par tous les n-grammes h^-w tels que hw appartient à l'ensemble des n-grammes ayant été observés :

$$\beta(h) = \frac{1 - \sum_{hw \in E} P^*[w|h]}{1 - \sum_{hw \in E} P^*[w|h^-]} \ . \tag{2.7}$$

Il est intéressant de noter que cette expression permet de conserver la propriété de somme à 1 des probabilités conditionnelles de chaque historique et des probabilités jointes sur l'ensemble des n-grammes théoriquement possibles (*cf.* Annexe A.1 pour s'en convaincre).

Les modèles de langue n-grammes fondés sur une telle méthode de repli sont les plus courants dans les systèmes de reconnaissance actuels. Toutefois, de nombreuses variantes visant à pallier certaines faiblesses de ces modèles sont régulièrement proposées.

2.1.2 Autres modèles de langue statistiques

Le modèle n-gramme de mots étant le modèle historique, de nombreuses variantes en ont été proposées. Celles-ci visent principalement à résoudre le fait que, les n-grammes de mots n'étant finalement que des séquences de quelques symboles contigus, ils n'intègrent aucune sémantique explicite et ne tirent pas profit de la nature structurée du langage naturel. Nous présentons ici les tentatives de modélisation qui reprennent en partie le principe du modèle n-gramme mais qui, en outre, cherchent à intégrer explicitement des connaissances linguistiques.

2.1.2.1 Modèles de langue *n*-classes

Les modèles de langue n-classes (*class-based language models*) proposent de regrouper les mots du vocabulaire en classes lexicales et de considérer principalement le calcul de la probabilité d'une séquence de mots comme celui de la probabilité d'une séquence de classes lexicales (Brown *et al.*, 1992). Dans le cas simple où chaque mot n'appartient qu'à une classe lexicale, la probabilité d'un mot sachant son historique s'écrit alors :

$$\begin{aligned} P[w_i|w_1...w_{i-1}] &= P[w_i, c_i|c_1...c_{i-1}] \\ &\approx P[w_i|c_i] \times P[c_i|c_1...c_{i-1}] \ , \end{aligned} \tag{2.8}$$

où c_i est la classe lexicale du mot w_i. En modélisant $P[c_i|c_1...c_{i-1}]$ par un modèle n-gramme de classes, la formule (2.8) s'écrit finalement :

$$P[w_i|w_1...w_{i-1}] = P[w_i|c_i] \times P[c_i|\mathbf{c_{i-n+1}}...c_{i-1}] \ . \tag{2.9}$$

Une telle stratégie de regroupement en classes a l'intérêt, d'une part, de diminuer l'éparsité des données en abaissant le nombre de paramètres à estimer pour le modèle – ce qui permet en contrepartie d'augmenter l'ordre du modèle – et, d'autre part, de permettre l'intégration d'informations linguistiques dans les modèles de langue en effectuant la classification des mots en classes lexicales d'après des propriétés linguistiques (Samuelsson et Reichl, 1999 ; Erdoğan, 2001). Dans une majorité de cas toutefois, cette classification s'effectue selon une méthode non supervisée fondée, par exemple, sur l'information mutuelle entre mots (Kneser et Ney, 1993 ; Martin *et al.*, 1998) ou sur des techniques d'analyse sémantique latente (Federico, 1996).

En pratique, le modèle n-classe est généralement interpolé avec un modèle n-gramme (Justo et Torres, 2007). Une généralisation de ce modèle peut consister à construire plusieurs niveaux hiérarchiques de classes parmi lesquels le niveau le moins abstrait réfère aux mots eux-mêmes (Dupont et Rosenfeld, 1997 ; Zitouni et Zhou, 2008). Dans ce cas, la probabilité d'une séquence de mots s'écrit comme l'interpolation des probabilités des classes des différents niveaux de la hiérarchie ou comme les probabilités de n-grammes mélangeant mots et classes (Mou *et al.*, 2002). Cette généralisation présente l'intérêt de permettre une technique de repli plus performante en proposant, dans le cas d'un n-gramme non observé, de se replier non pas directement sur un historique de longueur inférieure mais sur un historique correspondant à un niveau d'abstraction supérieur d'après la hiérarchie de classes (Zitouni, 2007).

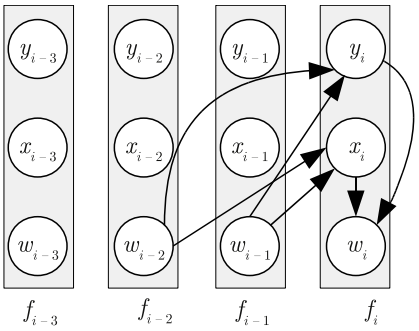


FIGURE 2.1 – Graphe de dépendance des paramètres d'une séquences de vecteurs. Un arc de A vers B signifie que B dépend de A (exemple tiré de (Kirchhoff *et al.*, 2008)).

2.1.2.2 Modèles de langue factoriels

Au sein d'un modèle de langue factoriel (*factored language model*), un mot n'est plus vu comme une simple chaîne graphique mais comme un vecteur de caractéristiques (Bilmes et Kirchhoff, 2003). En cela, il s'agit d'une généralisation des modèles de langue n-classes qui, à chaque un mot, associait comme unique caractéristique une ou plusieurs classes. Reprenant le principe des modèles n-grammes, la probabilité conditionnelle d'un mot sachant son historique s'écrit alors :

$$\begin{aligned} P[w_i|w_1...w_{i-1}] &= P[f_i|f_1...f_{i-1}] \\ &\approx P[f_i|f_{\mathbf{i-n+1}}...f_{i-1}] \ , \end{aligned} \tag{2.10}$$

où f_i est le vecteur de caractéristiques du mot w_i. Ces caractéristiques peuvent inclure le lemme et la classe grammaticale d'un mot, des morphes[8], son genre, son nombre, ou encore des booléens indiquant l'appartenance du mot à des classes sémantiques données...

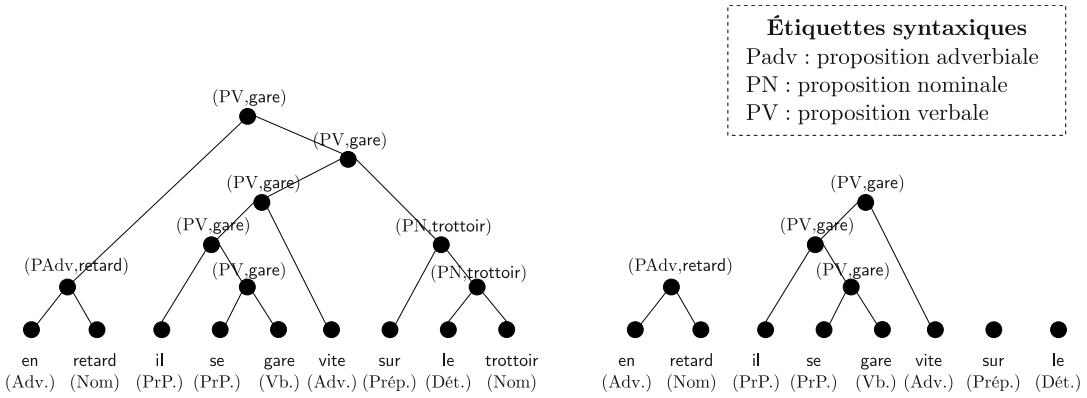
Pour limiter le problème d'éparsité des données qu'impliquerait l'apprentissage direct de probabilités de séquences de vecteurs[9], il est nécessaire de décomposer le calcul de $P[f_i|f_{i-n+1}...f_{i-1}]$ par un modèle de dépendances entre les caractéristiques des vecteurs. Par exemple, la figure 2.1 présente, pour le cas d'un modèle quadrigramme de vecteurs à 3 composantes, un graphe de dépendance qui conduit à écrire :

$$P[f_i|f_{i-3}f_{i-2}f_{i-1}] = P[w_i|x_i,y_i] \times P[x_i|w_{i-2},w_{i-1}] \times P[y_i|w_{i-2},w_{i-1}] \ . \tag{2.11}$$

De la même manière, les modèles factoriels peuvent utiliser des stratégies de repli variées. Pour cela, un graphe de repli définit quels paramètres d'un historique doivent être laissés de côté au moment d'un repli. Les modèles factoriels ont déjà montré de bons résultats sur le plan théorique (Smaili *et al.*, 2004) ou pour certaines tâches telles que la traduction automatique (Kirchhoff et Yang, 2005) ou l'étiquetage de séquences (Ji et Bilmes, 2005), mais, à notre connaissance, seule une étude s'est penchée sur leur utilisation

[8]Les morphes sont les plus petits signes linguistiques porteurs de sens qui composent les mots. Par exemple, le mot chanteur comporte les morphes chanter qui désigne l'action de chanter et eur qui désigne une personne qui fait cette action. En français, un mot se décomposent rarement en de nombreux morphes mais certaines langues, comme le turc ou le japonais, ont un fonctionnement tel que leurs mots rassemblent fréquemment 4 ou 5 morphes.

[9]Plus les vecteurs de caractéristiques sont complexes, plus rare devient chaque séquence de vecteurs dans le corpus d'apprentissage et moins fiable devient l'estimation des probabilités.



(a) Arbre syntaxique binaire complet (b) Analyse syntaxique partielle

FIGURE 2.2 – Représentation sous forme d'arbres binaires d'une analyse syntaxique complète (a) et partielle (b) de la séquence de mots « en retard il se gare vite sur le trottoir ».

en reconnaissance de la parole (Vergyri *et al.*, 2004). Toutefois, ce travail montre d'une part des gains très limités par rapport à l'utilisation d'un modèle n-gramme et s'intéresse d'autre part à une langue très flexionnelle, en l'occurrence l'arabe, c'est-à-dire une langue où les mots sont fléchis différemment selon leur fonction dans la phrase, ce qui n'est pas le cas de langues telles que le français ou l'anglais. L'intérêt des modèles factoriels pour la reconnaissance automatique de ces dernières langues reste donc encore à démontrer.

2.1.2.3 Modèles de langue structurés

Le modèle de langue structuré (*structured language model*) propose quant à lui de prendre en compte la structure syntaxique du langage naturel dans le calcul des probabilités conditionnelles (Chelba et Jelinek, 2000). Pour cela, ce modèle calcule les probabilités de couples (W, T) où W est une séquence de mots et T est une analyse syntaxique de W représentée sous la forme d'un arbre binaire. Les feuilles d'un tel arbre sont les mots de W accompagnés de leur classe morphosyntaxique et chaque nœud correspond à un syntagme[10] représenté par sa catégorie et sa tête (figure 2.2.a). Dans (Chelba *et al.*, 1997 ; Chelba et Jelinek, 2000), les auteurs définissent alors le calcul d'une probabilité conditionnelle comme :

$$P[w_i|w_1...w_{i-1}] = \frac{1}{Z(w_1...w_{i-1})} \sum_{T \in \mathcal{T}_1^{i-1}} \rho(w_1...w_{i-1}, T) \times P[w_i|w_1...w_{i-1}, T] \ , \qquad (2.12)$$

où $\mathcal{T}_1^{i-1}$ est l'ensemble des analyses syntaxiques partielles possibles à partir de la séquence $w_1...w_{i-1}$, T est l'un de ces arbres (*cf.* figure 2.2.b) et $\rho(w_1...w_{i-1}, T)$ est la probabilité que la séquence ait T pour analyse partielle. $Z(w_1...w_{i+1})$ est un facteur de normalisation.

En pratique, chaque analyse syntaxique partielle T peut être représentée comme la liste des racines de ses arbres syntaxiques. Dans l'exemple de la figure 2.2.b, la séquence de mots « en retard il se gare vite sur le » peut ainsi s'exprimer comme la séquence $<$(PAdv,retard) ;

[10] On appelle *syntagme* un ensemble de mots syntaxiquement valide et formant une unité sémantique. La taille d'un syntagme est comprise entre le mot et la phrase.

(PV,gare) ; (Prép,sur) ; (Dét,le)>. Si l'on note $r_1...r_{j(T)}$ cette séquence de racines, les calculs peuvent se simplifier en appliquant deux réductions markoviennes : l'une à ordre $n-1$ sur l'historique des mots et l'autre à l'ordre $k-1$ sur l'historique des racines :

$$P[w_i|w_1...w_{i-1},T] = P[w_i|h,h_T], \tag{2.13}$$
$$\text{avec } h = w_{\mathbf{i-n+1}}...w_{i-1} \text{ et } h_T = r_{\mathbf{j(T)-k+1}}...r_{j(T)} \ .$$

Les modèles de langue structurés reposent toutefois sur la fiabilité de l'analyseur syntaxique utilisé. Étant donné alors qu'aucune méthode satisfaisante n'existe pour l'instant pour effectuer l'analyse syntaxique d'un graphe de mots, l'utilisation d'un modèle de langue structuré au cours du processus de transcription semble encore difficilement envisageable. Pour cette raison, les modèles structurés sont appliqués à des listes de $\mathcal{N}$ meilleures hypothèses (Xu *et al.*, 2002), sur lesquelles des analyses syntaxiques fiables sont cette fois envisageables.

2.1.2.4 Modèles de langue exponentiels

Les modèles de langue exponentiels (*exponential language models*) représentent une famille de modèles dont le mode d'apprentissage ne repose pas sur le maximum de vraisemblance mais sur la recherche d'une distribution de probabilités d'entropie maximale (Rosenfeld, 1994)[11]. Ces probabilités s'expriment sous la forme générale suivante :

$$P[w_i|w_1...w_{i-1}] = \frac{1}{Z(h)} \exp\left[\sum_j \lambda_j \times f_j(h,w_i)\right], \tag{2.14}$$

où h correspond à la projection de l'historique $w_1...w_{i-1}$ vers sa classe d'équivalence, Z est un facteur de normalisation et chaque fonction f_j est une fonction indicatrice qui établit si, oui ou non, (h,w_i) présente la caractéristique j définie arbitrairement[12]. Les valeurs λ_j sont des paramètres à optimiser de telle sorte que le modèle de langue ait une entropie maximale et que, pour chaque j, la contrainte suivante soit respectée :

$$\sum_{h,w} P[h,w] \times f_j(h,w) = K_j \ , \tag{2.15}$$

où K_j est une masse de probabilité, généralement estimée à partir d'un corpus d'apprentissage. Il a été montré que, sous ces contraintes, la recherche des λ_j optimaux pouvait s'effectuer par des algorithmes itératifs (Darroch et Ratcliff, 1972 ; Della Pietra *et al.*, 1997).

Tout comme pour les approches basées sur le maximum de vraisemblance, le choix d'un historique est laissé libre. Il peut ainsi se limiter à un n-gramme de mots ou inclure des caractéristiques sémantiques ou syntaxiques indiquant, par exemple, la présence ou l'absence de certains mots importants dans l'historique (Lau *et al.*, 1993) ou encore la présence de patrons syntaxiques particuliers (Rosenfeld *et al.*, 2001). La puissance des

[11] En cela, il ne s'agit pas d'une alternative aux modélisations vues précédemment mais bien d'une autre manière d'effectuer l'apprentissage des modèles. Il est par exemple parfaitement possible de construire un modèle n-gramme exponentiel.

[12] Par exemple, est-ce que (h,w_i) est égal à une séquence bien précise ou est-ce que cette séquence présente une propriété linguistique donnée ?

modèles exponentiels tient alors dans le choix des fonctions f_j qui permettent de créer des dépendances entre caractéristiques de l'historique. Par ailleurs, ces modèles ne nécessitent pas explicitement de lissage puisque celui-ci est implicitement effectué par l'optimisation globale de la distribution[13]. Néanmoins, les modèles exponentiels sont connus pour leur temps de calcul élevé lié à la recherche des paramètres λ_j et au fait que l'approche par maximum d'entropie ne permet en réalité pas de calculer directement des probabilités conditionnelles mais des probabilités jointes, ce qui rend obligatoire le calcul de coûteux coefficients de normalisation (Rosenfeld, 1994).

2.1.2.5 Autres modèles

Nous avons jusqu'à présent détaillé les modèles ou types de modèles qui nous semblent les plus intéressants pour notre problème. Il ne faut toutefois pas perdre de vue que bien d'autres approches sont proposées dans la littérature. Dans la famille des modèles n-grammes, certains travaux suggèrent d'utiliser des historiques de longueurs variables (Bimbot *et al.*, 1995 ; Bonafonte et Mariño, 1996 ; Niesler et Woodland, 1996), de sauter ou d'inverser certains mots de l'historique (Rosenfeld, 1994 ; Langlois *et al.*, 2000) ou de mélanger ces différentes modélisations (Langlois *et al.*, 2003 ; Brun *et al.*, 2007). D'autres travaux récents proposent de construire la fonction de classification Φ sur des réseaux de neurones (Bengio *et al.*, 2001 ; Schwenk et Gauvain, 2002 ; Roark *et al.*, 2007). Ces techniques ont montré des gains de reconnaissance intéressants pour la transcription de parole conversationnelle par rapport à l'utilisation de modèles n-grammes (Schwenk et Gauvain, 2004). Ce regroupement en classes d'équivalence peut encore être guidé par la construction de forêts aléatoires d'arbres de décision à partir d'une représentation arbitraire des historiques de mots (Xu et Jelinek, 2004).

Indépendamment de leurs différences, toutes ces méthodes construisent des distributions de probabilités conditionnelles sur les mots d'un vocabulaire fini. Ceci permet de poser un cadre d'évaluation commun qui, *via* des mesures indépendantes de toute application concrète, permet de comparer la qualité de la modélisation linguistique de deux modèles de langue.

2.2 Évaluation d'un modèle de langue : mesure de la perplexité

Évaluer un modèle de langue statistique sur les sorties d'un système de reconnaissance automatique de la parole ne cible pas l'évaluation de la seule qualité de la modélisation linguistique étant donné que le processus de transcription dépend de nombreux autres facteurs, notamment de nature acoustique. Par ailleurs, le calcul d'un taux d'erreur est une opération coûteuse car il nécessite de lancer un processus de transcription complet sur un ensemble de test. D'autres mesures, indépendantes de tout cadre applicatif, sont alors utilisées pour évaluer la qualité du seul modèle de langue. Nous présentons la plus utilisée d'entre elles : la *perplexité*.

Le principe de la perplexité est de vérifier à quel point un modèle de langue permet de prédire les séquences de mots du langage qu'il est censé modéliser. En pratique, ces sé-

[13] Un lissage peut toutefois être nécessaire au moment d'estimer les masses de probabilité K_j de la formule (2.15).

quences de mots sont échantillonnées sous la forme d'un texte T représentatif du langage[14], et le pouvoir prédictif d'un modèle de langue M est calculé comme la log-vraisemblance moyenne de T sachant M :

$$L(T|M) = \frac{1}{n} \times \sum_{w_i \in T} \log_2 P_M[w_i|h_i] \ , \tag{2.16}$$

où h_i est l'historique du mot w_i dans la représentation choisie et P_M la probabilité conditionnelle que leur donne le modèle M. Notons que si l'un des mots w_i n'est pas dans le vocabulaire du modèle de langue, le calcul de $L(T|M)$ n'est pas possible. Dans ce cas, le mot n'est généralement pas compté et l'historique suivant, h_{i+1}, est calculé d'après la stratégie de repli définie par le modèle de langue. La perplexité du texte T d'après le modèle M se définit alors comme :

$$\mathrm{PPL}_M(T) = 2^{-L(T|M)} \ , \tag{2.17}$$

cette valeur étant d'autant plus faible que le modèle prédit bien les mots de T.

La perplexité peut être interprétée comme un « facteur de branchement » du modèle de langue, c'est-à-dire le nombre moyen de choix qui s'offrent au modèle pour chaque position i dans le texte pour trouver le bon mot w_i. La perplexité trouve également une interprétation dans le domaine de la théorie de l'information où elle est vue comme une fonction monotone de l'entropie croisée de la distribution estimée P_M avec la distribution théorique optimale du langage (*cf.* l'annexe A.2).

La perplexité est une mesure très communément utilisée de par la simplicité de son calcul. Toutefois, quelques précisions sont à apporter quant à son interprétabilité. Tout d'abord, de manière évidente d'après (2.16), la comparaison de deux perplexités n'a de sens que si celles-ci ont été calculées sur le même texte. Ensuite, il est préférable que les modèles de langue utilisés partagent un même vocabulaire. D'une part, ceci s'explique par le fait que, si certains mots du texte sont absents du vocabulaire d'un premier modèle mais sont présents dans celui d'un second modèle, les perplexités respectives de ces deux modèles ne seront pas calculées sur les mêmes ensembles de mots. D'autre part, d'après la contrainte de somme à 1 d'une distribution de probabilités, plus le vocabulaire d'un modèle est étendu, plus la valeur moyenne d'une probabilité fournie par ce modèle sera faible. Ainsi, la perplexité d'un modèle de langue aura tendance à être d'autant plus élevée que son vocabulaire est étendu. Enfin, dans le cadre de la reconnaissance automatique de la parole, la perplexité ne doit pas être vue comme une valeur corrélée au taux de reconnaissance d'un système. En effet, même si généralement une diminution de la perplexité se traduit par une diminution du WER, il existe de nombreux contre-exemples dans la littérature (Clarkson et Robinson, 1999). Il est toutefois plus vaguement accepté qu'une diminution de la perplexité inférieure à 5 % n'aura aucune incidence sur le WER, qu'une diminution comprise entre 10 et 20 % devrait se traduire par une diminution du WER et qu'une diminution supérieure à 30 % est très significative (Rosenfeld, 2000).

Après avoir donné des éléments sur la théorie des modèles de langue et sur la manière de les évaluer, la prochaine section aborde la question de l'apprentissage concret du vocabulaire et du modèle de langue n-gramme d'un système de reconnaissance automatique de la parole.

[14] Bien évidemment, ce texte ne doit pas faire partie du corpus d'apprentissage utilisé pour l'estimation du modèle de langue.

2.3 Procédure d'apprentissage

L'apprentissage d'un modèle de langue n-gramme pour un système de reconnaissance automatique de la parole est un processus fastidieux qui, à partir d'un ou plusieurs corpora d'apprentissage, consiste à sélectionner un vocabulaire à partir duquel est construit un lexique phonétisé, puis à estimer les probabilités n-grammes du modèle de langue. Cette section décrit ce processus à travers l'évocation des différentes étapes-clés que résume la figure 2.3. Nous en profitons également pour donner quelques précisions sur l'apprentissage des modèles de langue utilisés au sein du système IRENE.

2.3.1 Corpora d'apprentissage

Au delà du vocabulaire et du modèle de langue, la modélisation statistique du langage s'appuie avant tout sur le recours à des données textuelles d'apprentissage. La nature de ces données est primordiale dans le sens où celles-ci doivent être représentative de la tâche pour laquelle le vocabulaire et le modèle de langue seront utilisés. Il a notamment été montré que, pour une tâche donnée, il est préférable d'utiliser peu de données représentatives de cette tâche plutôt que beaucoup de données provenant d'une autre tâche (Rosenfeld, 2000 ; Lefevre *et al.*, 2001). Les données propres à une tâche étant toutefois généralement rares, un compromis consiste souvent à mélanger un petit corpus spécifique avec un large corpus textuel décorrélé de la tâche visée.

Le système IRENE étant à l'origine dédié à la transcription de journaux d'actualités, l'apprentissage de ses composants linguistiques s'est appuyé sur deux corpora : un premier corpus de 1 million de mots issus de transcriptions manuelles de journaux d'actualités radiodiffusés, corpus provenant des données d'apprentissage de la campagne ESTER (Galliano *et al.*, 2005) ; et un second corpus de 350 millions de mots constitués d'articles des archives du journal *Le Monde* entre 1987 et 2002.

2.3.2 Construction d'un vocabulaire et d'un lexique phonétisé

La construction d'un vocabulaire comprend principalement deux phases : la normalisation des textes des corpora d'apprentissage et la sélection des mots du vocabulaire. Dans le cadre d'un système de transcription, une troisième phase s'ajoute à celles-ci : la construction du lexique phonétisé.

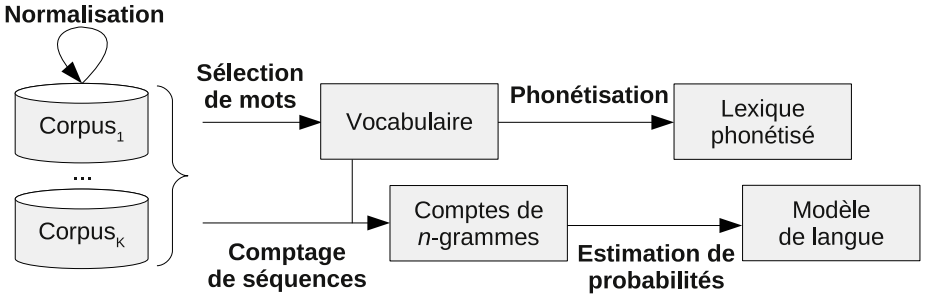


FIGURE 2.3 – Étapes importantes du processus d'apprentissage des composants linguistiques d'un système de reconnaissance automatique de la parole.

Normalisation. La normalisation des corpora d'apprentissage vise à homogénéiser l'ensemble des mots observés en transposant ceux-ci vers des formes standards. Ces dernières définissent les unités lexicales sur lesquelles sera construit le modèle de langue et qui, par la même occasion, composeront les sorties du système de transcription. Cette étape est un mal nécessaire car les textes bruts des différents corpora d'apprentissage sont très généralement issus d'auteurs différents avec des pratiques typographiques tout aussi différentes. Ces homogénéisations consistent alors à projeter des ensembles de mots ou séquences de mots jugés ambigus vers des formes de référence. En pratique, cela passe par l'établissement de règles de transformation *ad hoc* ou rendues plus ou moins génériques *via* l'usage intensif d'expressions régulières (table 2.1). Tout d'abord, cette normalisation des textes inclut généralement la rectification de l'orthographe de certains mots et la réécriture sous forme littérale, ou sous la forme transcrite de l'oral, de mots particuliers tels que les nombres, les abbréviations (titres de civilités, unités de mesures...) ou encore les adresses mél et autres url. Cette étape consiste également à définir la nature des unités lexicales en découpant ou regroupant certains mots – on parle souvent de *tokenisation*. Ces choix relèvent de l'intuition des personnes en charge de la construction des composants linguistiques quant à l'effet d'une tokenisation par rapport à une autre en terme de possibilité et de qualité de la modélisation du langage. Par exemple, il paraît linguistiquement intéressant de séparer les éléments de certains mots composés, tels que « -clé » dans l'exemple, de manière à permettre au modèle de langue de généraliser implicitement les mécanismes de construction de mots composés. À l'inverse, il peut être intéressant de considérer certaines séquences de mots comme une seule unité lexicale. Par exemple, il est préférable de considérer la séquence « il␣n'␣y␣a » comme un seul mot car, dans le cas contraire, cette séquence concentrerait sans doute toute la masse de probabilité de l'historique « il␣n'␣y », ce qui rendrait probablement impossible la reconnaissance de toute autre séquence ayant le même historique. Notons qu'il est toutefois parfois surprenant (et frustrant) de constater que certaines de ces intuitions, notamment guidées par des critères linguistiques, dégradent finalement la qualité des transcriptions automatiques. Enfin, la normalisation consiste également à découper les phrases des corpora d'apprentissage en groupes de souffles, unités que cherchent à modéliser les modèles de langue. Ainsi, toutes les phrases sont séparées les unes des autres et tous les signes de ponctuation sont supprimés. Finalement, les corpora d'apprentissage se présentent donc après normalisation comme un ensemble de séquences indépendantes d'unités lexicales séparées par des espaces. Dans la suite, par abus de langage, nous ferons référence à ces unités sous le terme de *mots*.

Sélection du vocabulaire. Fondée sur des corpora normalisés, la deuxième étape de la construction d'un vocabulaire consiste à sélectionner les quelques dizaines de milliers de mots sur la base desquels le modèle de langue sera appris. Un vocabulaire est construit de telle sorte qu'il soit optimal pour la tâche à laquelle sera dévolue le système de reconnaissance. Généralement, cette optimalité s'exprime à travers la notion de *couverture lexicale* du vocabulaire par rapport à un texte représentatif de la tâche en question. Cette couverture se définit comme le rapport entre le nombre d'occurrences des mots du vocabulaire V dans le texte T et le nombre total d'occurrences de tous les mots de ce même texte :

$$\text{couv}_V(T) = \frac{Card\{w \in T | w \in V\}}{Card\{w \in T\}} \ . \tag{2.18}$$

Patron lexical	→	Forme projetée
[aA]l- ?␣ ?Qu ?a[eiï]da	→	Al-Qaïda
[Mm]a ?c␣ ?[Dd]onald' ?s	→	McDonald's
clef	→	clé
m^2	→	mètre(s) carré(s)
Mme	→	Madame
anti- ?missiles	→	anti-␣missiles
accord- ?␣ ?clé	→	accord␣-clé
a␣priori	→	a_priori
il␣n'␣y␣a	→	il_n'_y_a

TABLE 2.1 – Exemples de règles de normalisation.

Le dual de cette notion, le *taux de mots hors vocabulaire*, est également souvent utilisé. Ce taux est fonction du nombre d'occurrences de mots du texte T qui ne sont pas dans le vocabulaire V et se calcule comme $1 - \text{couv}_V(T)$. Dans le cadre de la transcription de la parole, cette notion de couverture lexicale a un impact particulièrement fort sur les performances d'un système car chaque occurrence d'un mot absent du vocabulaire dans un document à transcrire se traduit par une voire plusieurs erreurs de transcription (Rosenfeld, 1995 ; Hetherington, 1995 ; Bazzi, 2002). Dans le système IRENE, le vocabulaire est ainsi construit en incluant tous les mots du corpus de transcription de journaux d'actualités radio-diffusés. Ce vocabulaire est ensuite complété par les mots les plus fréquents dans le corpus du journal *Le Monde* de manière à obtenir un vocabulaire de 65 000 mots. Notons que cette technique est très simple mais qu'il existe des approches plus sophistiquées pour construire un vocabulaire à partir de plusieurs corpora (Allauzen, 2003).

Phonétisation. Finalement, dans le cadre d'un système de reconnaissance automatique de la parole, la construction d'un lexique phonétisé à partir du vocabulaire sélectionné est nécessaire pour le bon déroulement du processus de transcription. Cette étape consiste à associer à chaque mot du vocabulaire une ou plusieurs phonétisations possibles. Cette phonétisation s'effectue principalement en tirant parti de ressources lexicales construites par des experts et en faisant appel à des outils de phonétisation automatique tels que LIAPHON (Béchet, 2001) ou IRISAPHON (Claveau, 2009). Il s'avère toutefois nécessaire de compléter cette partie automatique par un travail pénible de vérification et de phonétisation manuelle. Ce travail est notamment quasi nécessaire sur toutes les entités nommées et mots d'origine étrangère car ceux-ci sont généralement absents des dictionnaires phonétiques construits manuellement et sont encore mal traités par des systèmes de phonétisation automatique.

2.3.3 Estimation d'un modèle de langue *n*-gramme

Tel que nous l'avons vu à la section 2.1, un modèle de langue n-gramme est généralement estimé par une technique de maximum de vraisemblance. C'est notamment le cas des modèles de langue que nous utilisons au sein du système IRENE. Dans cette section, nous ne présentons donc que cette technique et laissons de côté l'apprentissage par maximum d'entropie. La chaîne de traitements nécessaire à un apprentissage par maximum de vrai-

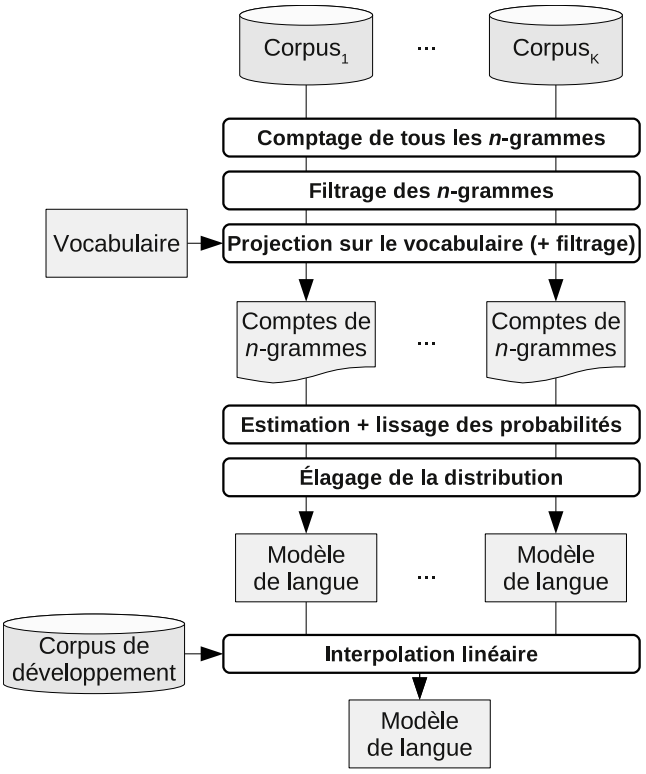


FIGURE 2.4 – Apprentissage d'un modèle de langue n-gramme.

semblance est schématisée par la figure 2.4. Elle consiste à compter les séquences de mots apparaissant dans les corpora d'apprentissage normalisés avant d'estimer des probabilités n-grammes.

Comptage de séquences. Pour chaque corpus d'apprentissage, le nombre d'occurrences de chaque séquence de 1 à n mots est compté et stocké. Parmi ces séquences, seules sont alors retenues celles ayant été vues plus d'un certain nombre de fois. Ce filtrage permet, d'une part, d'anticiper les problèmes d'estimation des probabilités des séquences rares et, d'autre part, de réduire à terme la taille du modèle de langue. Cette éviction de certains n-grammes est généralement appelée *cut-off* et le seuil utilisé pour celle-ci est référé par le terme *valeur de cut-off*. Les mots de chaque séquence restante sont ensuite parcourus et l'appartenance de chacun d'entre eux au vocabulaire est vérifiée – on parle de *projection* des séquences sur le vocabulaire. Lorsqu'une séquence contient un mot absent du vocabulaire, deux options sont alors possibles : soit cette séquence est supprimée, soit chaque mot hors vocabulaire est remplacé par une étiquette particulière. Cette dernière peut être commune à tous les mots hors vocabulaire ou dépendre de propriétés linguistiques de chaque mot remplacé, cette stratégie ouvrant la voie à des méthodes d'adaptation du vocabulaire que nous évoquons au chapitre 3. Lors de l'apprentissage de nos modèles de langue, chaque séquence contenant un mot hors vocabulaire a été supprimée.

Estimation de probabilités *n*-grammes. À partir de cette liste finale de fréquences de séquences, les probabilités sont estimées par maximum de vraisemblance (formule (2.2))

puis lissées, dans notre cas en utilisant un lissage de Kneser-Ney modifié (James, 2000). Enfin, un dernier élagage est effectué de manière à réduire la taille des modèle de langue, c'est-à-dire le nombre de n-grammes pour lesquels une probabilité conditionnelle est recensée. Cette élagage consiste à supprimer un ensemble des n-grammes de telle sorte que l'entropie de la distribution du modèle n'augmente pas plus que d'un certain pourcentage fixé. Finalement, un modèle de langue est ainsi obtenu pour chaque corpus d'apprentissage et la dernière étape consiste donc à fusionner ces différents modèles. Cette opération s'effectue sous la forme d'une interpolation linéaire des probabilités de chaque modèle de telle sorte que la distribution globale résultante minimise la perplexité d'un corpus de développement.

Au sein du système IRENE, les deux modèles de langue n-grammes utilisés, respectivement d'ordre 3 et 4, ont entièrement été appris selon le procédé que nous venons de décrire *via* la boîte à outils SRILM (Stolcke, 2002). Accessoirement, suite à ces différentes étapes, nos modèles ont en outre été compilés vers un format propre à notre décodeur SIROCCO.

Maintenant que nous avons décrit les principes théoriques et le processus d'apprentissage d'un vocabulaire et d'un modèle de langue n-gramme, nous possédons un recul suffisant pour souligner les limites de tels composants dans le cadre de la reconnaissance automatique de la parole, et plus particulièrement pour la transcription de documents audio présentant des thèmes variés.

2.4 Limites des modèles *n*-grammes

Parmi les différents types de modèles de langue que nous avons décrits, les modèles n-grammes sont ceux qui permettent d'atteindre les meilleures performances en reconnaissance. Pourtant, ces modèles présentent de nombreuses limites qui tiennent tant à leur nature qu'à leur processus d'apprentissage. Cette section présente quelques unes d'entre elles et souligne leur impact face aux changements de thèmes qui s'opèrent entre différents documents audio à transcrire ou au sein d'un même document.

2.4.1 Limites générales

De nombreuses critiques d'ordre général sont fréquemment adressées envers les modèles de langue n-grammes. Tout d'abord, l'hypothèse d'indépendance faite par la modélisation d'un historique sous la forme d'un n-gramme s'avère fausse en pratique. La lecture de n'importe quel texte permet de s'apercevoir qu'il existe des rapports de cause à effet entre des mots séparés par bien plus que ce que ne couvre le spectre d'un n-gramme. Ensuite, les modèles n-grammes ne modélisent rien d'autre que des séquences de symboles faisant office de mots. Cette représentation omet toutes les dimensions linguistiques de chaque mot. Ceci limite bien évidemment la portée des traitements qui peuvent être opérés lors du décodage pour s'assurer au plus tôt de la cohérence des transcriptions en cours de construction. Aussi, les développeurs de systèmes sont souvent contraints à travailler sur les listes de meilleures hypothèses à la sortie du décodeur et à utiliser des outils de traitement automatique des langues afin de réintégrer des informations de haut niveau aux séquences de mots. Enfin, comme nous l'avons vu à la section précédente, l'apprentissage d'un vocabulaire et d'un modèle de langue n-gramme suit un processus complexe et lourd[15]. Ces composants sont

[15] En comparaison avec une majeure partie des alternatives qui s'offrent au modèle n-gramme, cet apprentissage peut toutefois être compté parmi les plus simples.

ainsi généralement appris une fois pour toutes et n'évoluent donc pas en dépit des multiples variations du langage qui peuvent se faire sentir d'un document à transcrire à l'autre, voire au sein d'un même document. Ces variations sont généralement présentées autour de trois axes (Bellegarda, 2004). Premièrement, le langage évolue avec le temps. Par exemple, on n'emploie pas les mêmes mots ni les mêmes expressions aujourd'hui qu'il y a 20 ans. Deuxièmement, de la même manière qu'on parle différemment lors d'une réunion de travail ou lors d'une discussion avec un ami, l'utilisation du langage varie en fonction du genre d'un document. Enfin, d'un thème de discussion à l'autre, la nature et l'emploi du vocabulaire ne sont pas les mêmes. On n'utilise par exemple pas les mêmes mots ni séquences de mots pour parler de l'inflation du prix du pétrole ou pour parler de football. C'est justement le cas de ces variations thématiques qui nous intéresse dans cette thèse.

2.4.2 Le cas des variations thématiques

Dans le cadre de la transcription de journaux d'actualités, la variabilité des thèmes rencontrés est une caractéristique intrinsèque. Avant de définir l'impact linguistique de ces variabilités, nous nous attardons quelques instants sur la notion de thème.

2.4.2.1 Qu'est ce qu'un thème ?

Si, implicitement, un thème peut se définir comme « le sujet dont on parle », en donner une définition formelle est nettement plus difficile. Au sein des différentes communautés qui s'intéressent à la notion de thème, cette question n'est d'ailleurs pas tranchée.

En linguistique, le thème représente pour certains le socle commun d'informations partagées par les participants à une communication, notion s'opposant alors au *rhème*, information nouvelle qu'apporte la communication en supplément du thème (Marandin, 1988 ; Ghadessy, 1995). Une autre approche consiste à définir le thème d'un discours comme une *isotopie sémantique*, c'est-à-dire comme un ensemble de traits sémantiques, ou *sèmes*, stables et récurrents au sein du discours, ces traits sémantiques émanant des signifiés de chaque mot du discours (Rastier, 1995). Alternativement, d'autres travaux adoptent une définition différentielle en caractérisant les changements de thèmes au sein d'un discours plutôt que les thèmes eux-mêmes (Brown et Yule, 1983).

Parmi ces définitions, le traitement automatique des langues se focalise sur celles qui sont applicables automatiquement, c'est-à-dire sans faire appel au jugement humain. Au sein de ce domaine, trois tâches principales peuvent être dégagées des travaux qui ont trait à la notion de thème. Il s'agit soit de segmenter un long texte en portions thématiquement cohérentes, soit de caractériser le thème d'un document ou enfin de détecter la présence d'un thème donné au sein d'un texte. En reprenant l'idée que la présence d'un thème se définit souvent comme une isotopie sémantique et que, en pratique, les sens sont portés par les mots d'un discours, la notion de thème s'exprime alors souvent au sein de ces tâches comme un ensemble de mots présentant une forme de cohésion tout au long du discours et définissant, par leur co-présence, le thème même abordé. Certains travaux cherchent ainsi à déterminer l'ensemble de ces mots qui caractérisent le thème par la confrontation du vocabulaire de textes spécialisés avec un ensemble de textes généralistes (Drouin, 2004). D'autres travaux traitent le problème de la segmentation thématique en recherchant des frontières de part et d'autres desquelles la distribution des mots employés diffère (Hearst, 1997 ; Ferret, 1998). À l'inverse de ces efforts, certains posent la notion de thème comme un ensemble de sujets pré-déterminés que peut aborder un discours, le nombre de sujets prédéfinis pouvant varier

de quelques uns (sport, politique, science...) à plusieurs milliers (Seymore et Rosenfeld, 1997 ; Yamashita *et al.*, 1998 ; Brun, 2003).

Ces multiples définitions et interprétations laissent transparaître que la notion de thème reste floue mais qu'elle se caractérise avant tout par des propriétés de cohérence et de saillance sémantique des mots d'un discours. Ainsi, dans l'absolu, la granularité d'un thème ne devrait pas pouvoir être déterminée *a priori*. Par exemple, une longue émission sur la guerre en Irak pourrait probablement être découpée en sous-thèmes représentant chaque facette de ce conflit alors que, à l'inverse, la présence d'un bref volet sur ce même conflit au milieu d'une émission de divertissement pourra vraisemblablement être considérée comme l'évocation d'un seul et unique thème. Dans nos travaux, nous ne nous risquons à aucune définition de la notion de thème et ne nous appuyons sur aucune classification ou hiérarchisation des thèmes les uns par rapport aux autres. Plus précisément, nous nous appuyons dans nos travaux sur un ensemble de documents issus d'une segmentation thématique manuelle donnée de journaux d'actualités. Nous ne chercherons pas à juger ce découpage, en particulier, en ce qui concerne sa granularité.

2.4.2.2 Variations thématiques et modélisation linguistique

Au delà de la définition d'un thème, la question qui nous intéresse est de savoir en quoi la modélisation linguistique change ou doit changer lorsqu'il s'agit de transcrire les mots de textes traitant de thèmes variés.

La principale propriété d'un thème est qu'il réfère de manière récurrente à des notions qui lui sont spécifiques. Sur le plan linguistique, un thème se manifeste alors par l'apparition singulièrement fréquente des mots ou des séquences de mots qui portent ces notions spécifiques. Ainsi, les modélisations du langage en général, et du langage au sein d'un thème, se distinguent par des probabilités d'apparition différentes de ces mots et séquences de mots. Plus généralement, ces spécificités induisent également une modification des probabilités des séquences de mots au sein desquelles s'insèrent correctement les différentes « expressions thématiques » dans un discours. Cette « propagation » peut être aussi bien guidée par la syntaxe du langage général que par des spécificités d'emploi liées au thème considéré. Par exemple, si une émission traite de football, les mots et expressions comme ballon, frapper dans la balle ou siffler un coup franc seront bien plus probables que s'il s'agit d'une émission à propos d'un autre thème et, dans un discours, des séquences comme « un bon coup franc vient d'être sifflé » ou « il reprend le ballon du pied gauche » deviennent elles aussi plus probables.

Dans le cas d'une modélisation n-gramme telle que nous l'avons décrite à la section 2.3, ces spécificités lexicales et linguistiques posent différents problèmes. D'une part, nombre de mots spécifiques à un thème donné sont par défaut absents du vocabulaire initial car ceux-ci ont été jugés trop rares lors de l'apprentissage du système. Ces mots, typiquement des termes techniques ou entités nommées, doivent alors être ajoutés au vocabulaire et leurs usages linguistiques doivent être déterminés et valués sous la forme de probabilités n-grammes. D'autre part, les probabilités recensées par le modèle de langue ne reflètent que la moyenne des usages observés pour un très grand nombre de thèmes au sein du corpus d'apprentissage. Ainsi, certaines probabilités du modèle de langue doivent être adaptées pour coller au mieux à l'usage particulier de la langue au sein du thème considéré. Ce travail d'adaptation est d'autant plus difficile que, comme souligné précédemment, les modèles n-grammes n'intègrent aucune information linguistique sur les mots et que, outre

l'estimation de probabilités, la détermination des n-grammes à adapter se révèle alors être une tâche également ardue.

Au cours de ce chapitre, nous avons exposé le principe de la modélisation statistique du langage à travers différents exemples de modèles tirés de la littérature. Parmi ceux-là, le modèle n-gramme se présente comme le plus efficace. C'est pourquoi il est celui qu'intègrent le plus fréquemment les systèmes de transcription actuels, dont celui que nous utilisons. En mettant en parallèle le lourd processus d'apprentissage d'un tel modèle avec les variations linguistiques induites par des thèmes différents dans des documents à transcrire, il apparait comme nécessaire de réfléchir à une méthode d'adaptation d'un modèle de langue et de son vocabulaire à chaque thème rencontré dans un discours. Ce problème d'adaptation est bien connu dans le domaine de la reconnaissance automatique de la parole ; le prochain chapitre présente un état de l'art des travaux qui s'y intéressent.

Chapitre 3

Adaptation linguistique d'un système de reconnaissance automatique de la parole

Les systèmes actuels de reconnaissance automatique de la parole à grand vocabulaire sont amenés à traiter des documents audio de natures diverses. Pour ce faire, le vocabulaire et le modèle de langue n-gramme de ces systèmes sont appris une fois pour toute sur de vastes corpora de textes variés afin d'obtenir une connaissance générale et étendue de la langue. Ces composants manquent alors de précision pour modéliser les spécificités lexicales et linguistiques des différents types de documents que le système est amené à transcrire. L'adaptation linguistique d'un système de reconnaissance vise alors à modifier, sans relancer un lourd processus d'apprentissage, le vocabulaire et le modèle de langue d'un système pré-existant pour une tâche de transcription donnée. Cette tâche peut se caractériser par des documents audio à traiter datant d'une certaine époque, relevant d'un genre particulier[16] ou encore traitant d'un thème donné. De manière générique, si aucune précision n'est faite quant à ce qui distingue la tâche visée de celle pour laquelle a initialement été appris le système de transcription, on parle alors d'*adaptation à une tâche*. Au sein de cette problématique, nous nous intéressons plus spécifiquement au problème de l'*adaptation thématique*, problème qui vise à modifier les composants linguistiques du système en fonction du seul thème d'un document à transcrire, sans prêter attention à son époque ou à son genre. Cette problématique plus précise représente une proportion non négligeable des travaux d'adaptation de la littérature – bien qu'une partie d'entre eux reste relativement évasif quant à l'axe d'adaptation qui les intéresse. Cet intérêt se justifie notamment par la propension particulière qu'ont les thèmes à faire varier l'usage de la langue et par le fait que l'adaptation thématique s'inscrit dans un cadre applicatif en pleine émergence, à savoir la transcription de flux multimédia, et plus particulièrement d'émissions d'actualités, pour la création de nouveaux services télévisuels. Globalement toutefois, de nombreuses questions sont communes à la problématique d'adaptation à une tâche quelconque et à celle de l'adaptation thématique. Bien que nous nous intéressions donc précisément au problème de l'adaptation au thème d'un document, nous évoquons également dans ce chapitre les travaux qui s'intéressent à d'autres tâches.

[16] Par exemple, des conversations entre amis alors que le système de départ a été appris sur de la parole préparée et sur des textes écrits.

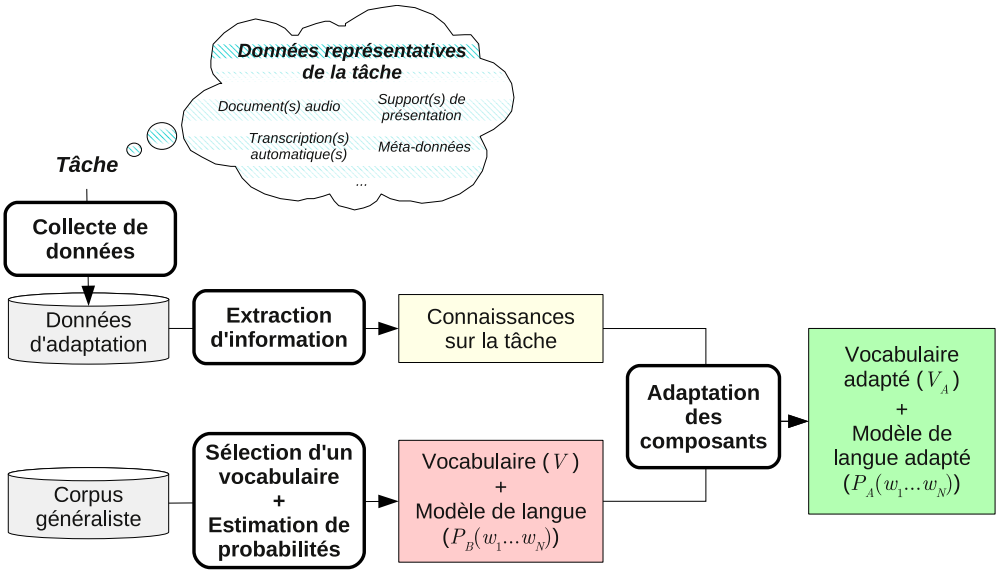


FIGURE 3.1 – Schéma général de l'adaptation linguistique d'un système de transcription.

De nombreuses techniques d'adaptation linguistique à une tâche ont été proposées dans la littérature. Dans l'ensemble, elles partagent un schéma de fonctionnement commun que présente la figure 3.1 inspirée de (Bellegarda, 2004). De manière synthétique, étant donné un système de reconnaissance appris sur un corpus généraliste, l'adaptation vise à recueillir des données propres à la tâche de transcription visée afin d'en extraire des connaissances spécifiques, puis à utiliser ces connaissances pour modifier le vocabulaire d'origine et réestimer les probabilités du modèle de langue généraliste. En pratique, pour amorcer ce processus, la tâche d'adaptation est décrite par des données textuelles qui en sont représentatives. Il peut par exemple s'agir de mots-clés ou de quelques textes de référence donnés par des experts pour décrire un thème (Sethy *et al.*, 2005). Plus généralement cependant, la tâche consiste à adapter le système en vue de transcrire au mieux un ou plusieurs documents audio. La caractérisation de la tâche d'adaptation s'effectue alors souvent en générant une première transcription automatique à partir du système généraliste. Dans certains cas, ces transcriptions automatiques peuvent être complétées par d'autres ressources. Par exemple, il est possible de caractériser la tâche *via* des supports de présentation accompagnant des enregistrements de réunions (Yamazaki *et al.*, 2007) ou des textes descriptifs (titre, synopsis...) accompagnant des émissions archivées (Allauzen et Gauvain, 2005b) ou diffusées sur Internet (Ogata *et al.*, 2007).

La réalisation de la chaîne d'adaptation implique de savoir répondre à de nombreuses questions. Premièrement, quel type de données d'adaptation doit être utilisé et comment récupérer ces données pour la tâche considérée ? Ensuite, quelles informations sont nécessaires pour l'adaptation du modèle de langue et comment effectuer cette adaptation en fonction de la nature des informations extraites ? Enfin, quels mots doivent être ajoutés au vocabulaire, comment peut-on trouver ces mots et comment peuvent-ils être pris en compte par le modèle de langue ? La littérature comporte beaucoup de travaux liés à ces différentes questions sans toutefois que tous s'intéressent à l'intégralité de la chaîne de traitements. De fait, une grande hétérogénéité existe parmi ces travaux : certains ne procèdent qu'à l'adaptation du modèle de langue ou du vocabulaire ; par ailleurs ces travaux

ne ciblent pas tous la même tâche d'adaptation et tous n'abordent pas non plus le problème avec la même philosophie, notamment quant au niveau de supervision toléré pour la mise en œuvre du processus d'adaptation. Cette section tente donc de dégager une vue d'ensemble de l'existant afin d'en exposer les grandes tendances autour des trois points que sont la récupération de données d'adaptation, l'adaptation du modèle de langue et celle du vocabulaire.

Bien que cette présentation vaille globalement pour une tâche d'adaptation quelconque, nous attachons une importance particulière à celle qui nous intéresse, l'adaptation thématique. Rappelons à ce propos que l'optique de nos travaux est de proposer une chaîne d'adaptation thématique complète et entièrement non supervisée, celle-ci devant notamment faire abstraction de toute connaissance *a priori* quant aux éventuels thèmes pour lesquels une adaptation devrait être lancée. Nous souhaitons en ce sens privilégier au maximum l'usage de techniques de traitement automatique des langues permettant de manipuler des éléments linguistiques explicites tels que des mots ou des séquences de mots. Par ailleurs, la méthode d'adaptation visée doit rester dans le cadre d'une modélisation n-gramme fondée sur un vocabulaire fini. Nous soulignons ainsi, lorsque cela s'avère nécessaire, les forces et les faiblesses des méthodes de l'état de l'art au regard de ces impératifs que nous nous sommes fixé pour le développement d'un processus d'adaptation thématique.

3.1 Récupération de données d'adaptation

La modélisation statistique du langage se heurte à un paradoxe : d'un côté, il faut beaucoup de données d'apprentissage pour constituer un vocabulaire avec une bonne couverture lexicale et estimer de manière fiable les probabilités du modèle de langue alors que, d'un autre coté, la qualité de ces composants pour une tâche donnée dépend très largement de l'adéquation entre le corpus d'apprentissage et la tâche considérée (Rosenfeld, 1996 ; Rosenfeld, 2000). En cela, le choix de données pertinentes est primordial pour la tâche d'adaptation d'un système de transcription. Ce choix s'articule autour de deux critères : la source au sein de laquelle les données d'adaptation sont recherchées, cette source influençant leur nature, et la manière dont elles sont sélectionnées.

3.1.1 Provenance et nature des données

Pour la récupération de données d'adaptation, deux types de ressources textuelles sont utilisés : des données hors-ligne et Internet.

Historiquement, l'apprentissage de modèles de langue pour la reconnaissance de la parole s'effectue principalement sur des documents écrits, typiquement des archives de journaux, ainsi que sur des transcriptions manuelles. Avec la numérisation des contenus et la création de campagnes d'évaluation de tâches sur les langues naturelles, ces corpora se sont largement développés. Citons par exemple, pour le français, les corpora écrits des journaux *Le Monde* et *Le Monde diplomatique*, ainsi que les corpus du français parlé ESTER et BREF, ou, pour l'anglais écrit et parlé, le *Corpus of Comtemporary American English*, le *British National Corpus* ou encore le corpus *Gigaword*. Pour la tâche d'adaptation du vocabulaire, ces ressources peuvent servir à construire des vocabulaires de repli (*fallback lexicon*) beaucoup plus riches que le vocabulaire effectivement pris en compte par le modèle de langue du système (Geutner *et al.*, 1998a ; Chung *et al.*, 2004 ; Palmer et Ostendorf, 2005 ; Ohtsuki *et al.*, 2005 ; Scharenborg *et al.*, 2007). Certains travaux proposent également

de tirer parti de ces très vastes bases textuelles hors-ligne pour y rechercher des textes susceptibles de correspondre à la tâche d'adaptation (Seymore et Rosenfeld, 1997 ; Klakow, 2000 ; Chen *et al.*, 2004). Malgré leur taille et leur disponibilité, ces bases ont toutefois deux inconvénients. D'une part, comme elles sont statiques et fermées, rien ne garantit d'y trouver des données d'adaptation réellement pertinentes. D'autre part, elles restent majoritairement constituées de textes écrits dont le style diffère du langage parlé.

Ainsi, beaucoup de techniques d'adaptation récentes proposent plutôt d'utiliser le Web comme un corpus ouvert (Berger et Miller, 1998 ; Kilgarriff et Grefenstette, 2003). Les ressources qui y sont recensées présentent l'intérêt d'être en permanence au fait de l'actualité et de traiter de quasi tous les thèmes possibles. Deux techniques de rapatriement des données du Web sont principalement rencontrées en fonction de la tâche visée. Certains travaux proposent de collecter régulièrement des données sur des sites d'actualités pour constituer ou mettre à jour une base textuelle hors-ligne (Kemp et Waibel, 1998 ; Auzanne *et al.*, 2000 ; Bertoldi et Federico, 2001 ; Allauzen et Gauvain, 2005a ; Martins *et al.*, 2006). Si ces techniques permettent effectivement de réduire significativement les taux d'erreurs de transcriptions d'émissions d'actualités, leur efficacité semble plus incertaine pour transcrire d'autres types de documents audio où les thèmes traités ne sont pas liés à des sujets d'actualités. D'autres travaux procèdent alors à la récupération de textes plus ciblés *via* l'utilisation de moteurs de recherche en ligne (*Google*, *Yahoo !*...) (Yu *et al.*, 2000 ; Nisimura *et al.*, 2001 ; Sethy *et al.*, 2005 ; Ogata *et al.*, 2007 ; Oger *et al.*, 2008 ; Aronowitz, 2009). La popularité des approches en ligne se justifie en partie par le fait que les textes issus d'Internet ont de meilleures propriétés que les textes écrits classiques pour modéliser l'oral et, notamment, la parole conversationnelle (Bulyko *et al.*, 2007). Par exemple, des pages Web contiennent en général des proportions plus élevées de pronoms personnels aux première et deuxième personnes que des textes écrits classiques (Vaufreydaz *et al.*, 1999). Cette qualité a néanmoins un prix puisque le contenu linguistique principal d'une page Web est généralement bruité par de nombreuses fautes d'orthographe et par d'autres contenus comme des menus, des publicités... Ce dernier point fait l'objet d'efforts de diverses communautés pour développer des outils de nettoyage de pages Web (Baroni *et al.*, 2008). Par ailleurs, d'autres données venant d'Internet sont également parfois utilisées pour constituer des données d'apprentissage, par exemple des emails ou des données venant de *newsgroups* (Vaufreydaz, 2002 ; Brun, 2003 ; Schwarm *et al.*, 2004). Bien que ces données soient moins bruitées et présentent un style généralement encore plus proche de la langue parlée comparativement à des pages Web, il s'agit de ressources difficilement collectables automatiquement.

Finalement, notons que certains travaux éludent l'étape de récupération de données d'adaptation en ne s'appuyant ni sur de vastes corpora hors-ligne ni sur des textes d'Internet mais en se référant directement aux données textuelles qui décrivent la tâche d'adaptation. Par exemple, dans (Souvignier *et al.*, 2000 ; Gretter et Riccardi, 2001), les probabilités d'un modèle de langue initial sont modifiées selon un principe qui s'apparente à du *relevance feedback* en tenant compte de mesures de confiance et probabilités postérieures provenant directement de listes des $\mathcal{N}$ meilleures hypothèses ou de réseaux de confusions du document à transcrire. Toutes ces ressources directement liées à la tâche d'adaptation sont toutefois souvent de taille réduite et difficilement exploitables directement, notamment pour la réestimation de probabilités. Elles sont ainsi généralement utilisées indirectement en guidant la sélection de données pertinentes à partir d'un corpus hors-ligne ou de pages Web.

3.1.2 Sélection des données

La sélection de documents représentatifs d'une tâche donnée dépend largement du cadre applicatif posé par cette tâche. Dans certains cas, la tâche est déterminée *a priori* et la sélection de données pertinentes peut ainsi se faire en toute connaissance de cause. Il peut notamment s'agir de développer rapidement, à partir d'un système généraliste pré-existant, un système de transcription destiné à une utilisation durable pour traiter des documents de même nature, par exemple des comptes-rendus post-opératoires. À l'inverse, d'autres applications visent à adapter régulièrement un système initial pour transcrire des documents variés, notamment sur le plan thématique, auquel cas la recherche de données d'adaptation s'avère plus difficile. Une distinction doit donc être faite entre stratégies supervisées, où la tâche est préalablement définie, et non supervisées.

Lorsque la tâche d'adaptation est clairement établie et n'est pas amenée à changer régulièrement, il peut être intéressant de s'appuyer sur des ressources textuelles spécifiques construites une fois pour toutes de manière supervisée. Par exemple, dans (Ariki *et al.*, 2003), les auteurs s'intéressent uniquement à la transcription de retransmissions sportives et utilisent directement des transcriptions manuelles de matchs pour adapter leur système de reconnaissance. Comme ces données sont néanmoins coûteuses à créer, elles peuvent être en quantité limitée et servir alors uniquement à amorcer un processus de récupération d'une plus large collection de textes (Sethy *et al.*, 2005). Dans d'autres cas, la supervision consiste à considérer un ensemble fini et constant de dimensions à partir desquelles une tâche peut être représentée et à déterminer ensuite où se place exactement chaque document à transcrire au sein de cet espace. Concrètement, dans le cas de l'adaptation thématique, cette pratique consiste à fixer *a priori* un certain nombre de thèmes figés pouvant être évoqués et à constituer pour chacun de ces thèmes une collection spécifique à partir d'un vaste corpus généraliste hors-ligne (Seymore et Rosenfeld, 1997 ; Chen *et al.*, 1998 ; Brun, 2003 ; Lane *et al.*, 2005). Notons que ce découpage s'appuie généralement sur un (coûteux) étiquetage thématique manuel des textes. Des techniques de *clustering* de textes permettent cependant de rendre cette approche plus générique (Carlson, 1996 ; Kneser et Peters, 1997). Dans tous les cas, la sélection de données d'adaptation s'apparente alors à un problème de catégorisation de textes (Sebastiani, 2002) où il faut associer les données descriptives de la tâche, par exemple une transcription automatique, à une ou plusieurs classes thématiques. Ces méthodes présentent l'inconvénient majeur de restreindre l'adaptation à un nombre, aussi grand soit-il, limité et statique de thèmes. Elles ne permettent notamment pas de prendre en compte le fait que les thèmes évoluent et que certains sujets nouveaux apparaissent avec le temps. Par ailleurs, le fait d'arrêter une unique classification des thèmes semble contradictoire avec le fait que la granularité du thème d'un document est variable en fonction du contexte thématique dans lequel se place ce document.

D'autres approches proposent alors de construire les collections textuelles d'adaptation de manière automatique, c'est-à-dire sans supposer aucune structure prédéfinie de la tâche. Si un corpus hors-ligne est disponible, ce travail consiste à parcourir un ensemble de textes et à déterminer lesquels d'entre eux coïncident le mieux avec la tâche visée. Dans le cas d'une adaptation thématique, des méthodes s'inspirent pour cette sélection du domaine de la recherche d'information. Les transcriptions automatiques de chaque document pour lequel l'adaptation du système doit être faite peuvent notamment être utilisées au sein d'un moteur de recherche hors-ligne pour récupérer des données d'adaptation (Kemp et Waibel, 1998 ; Klakow, 2000). L'élément-clé de ces méthodes est la définition d'une mesure de

similarité entre le document-requête et les textes de la base de référence. Aux travaux qui utilisent directement des techniques issues de la recherche d'information (Mahajan *et al.*, 1999 ; Niesler et Willett, 2002) s'ajoutent alors d'autres qui s'appuient, pour leur part, sur des principes de la théorie de l'information (Chen *et al.*, 2003 ; Bigi *et al.*, 2004).

Dans le cas où la ressource linguistique est Internet, la comparaison selon un critère personnalisé d'une transcription automatique avec des documents d'une base hors-ligne n'est pas directement envisageable et il est nécessaire de passer par des moteurs de recherche en ligne. La tâche de sélection des documents pertinents est alors laissée à ces moteurs indépendants et le problème devient plutôt de construire des requêtes pertinentes (Gao *et al.*, 2006 ; Suzuki *et al.*, 2006 ; Ito *et al.*, 2008). Toute la difficulté tient bien sûr dans le fait de savoir formuler ces requêtes : il faut à la fois qu'elles caractérisent bien la tâche mais qu'elles aboutissent aussi à la récupération de suffisamment de pages Web pour envisager l'emploi d'outils statistiques (Fouqueré, 2003). Dans le cadre d'une adaptation au genre, ces requêtes peuvent par exemple être des n-grammes fréquents comme « *I never thought I would* » ou « *I would think so* » (Bulyko *et al.*, 2003). Dans le cadre d'une adaptation thématique, il s'agit plus généralement de sélectionner des mots-clés à partir des données liées à un document traitant d'un thème unique. Tout comme pour l'utilisation de corpora hors-ligne, on retrouve des travaux qui s'inpirent du domaine de la recherche d'information pour extraire des mots-clés à la fois fréquents et discriminants (Suzuki *et al.*, 2006), par exemple en se basant sur la pondération *tf-idf* (Salton, 1989). Alternativement, dans (Sethy *et al.*, 2005), les auteurs mesurent l'entropie relative entre un modèle de langue généraliste et un modèle spécialisé appris à partir d'un petit corpus thématique initial pour trouver les n-grammes les plus caractéristiques du thème considéré. Bien que ces stratégies permettent d'exploiter la richesse d'Internet, il n'en demeure pas moins qu'elles laissent la sélection des documents pertinents aux bons soins des moteurs de recherche en ligne. Malgré les performances incontestables de ces derniers pour nos usages quotidiens, rien ne garantit toutefois que les résultats qu'ils retournent soient ceux qui nous intéressent pour la tâche d'adaptation. À notre connaissance, peu de travaux font pourtant l'effort de vérifier la qualité de ces résultats.

Après avoir abordé le problème de la collecte de données d'adaptation, voyons comment celles-ci sont utilisées pour la réestimation d'un modèle de langue.

3.2 Adaptation du modèle de langue

L'adaptation d'un modèle de langue vise à pallier à moindre frais le manque d'adéquation des modèles généralistes utilisés par défaut pour modéliser le langage spécifique à une tâche donnée. Cette adaptation cherche notamment à ne pas réitérer le processus d'apprentissage complet d'un modèle de langue et s'appuie pour cela sur des données représentatives de la tâche obtenues par l'un des processus que nous venons de décrire. La question est alors de savoir comment modifier les probabilités n-grammes du modèle de langue initial pour prendre en compte les caractéristiques linguistiques de ces données et finalement obtenir un modèle adapté conduisant à une perplexité et un WER les plus bas possibles pour la tâche visée. Nous notons tout au long de cette section P_B la distribution de probabilités du modèle initial, P_A celle du modèle adapté à déterminer et P_a celle d'un modèle de langue d'adaptation estimé à partir des seules données d'adaptation.

3.2.1 Interpolation de probabilités ou de comptes

Initialement, une majorité de travaux ont proposé de calculer la distribution du modèle adapté par la combinaison des probabilités du modèle de langue initial avec celles d'un plus petit modèle de langue estimé à partir des données d'adaptation. Plus généralement, ce problème se décrit comme la combinaison d'un nombre variable de distributions existantes. Sous sa plus simple expression, cette combinaison est opérée par une interpolation linéaire,

$$P_A[w|h] = \sum_i \lambda_i \times P_i[w|h] \ , \tag{3.1}$$

où les P_i sont les distributions de modèles de langue estimées à partir de ressources différentes et les λ_i sont des *coefficients d'interpolation* compris entre 0 et 1 sous la contrainte que $\sum_i \lambda_i = 1$. Ces coefficients peuvent être déterminés en minimisant la perplexité de P_A sur un corpus de développement, par exemple un sous-ensemble des données d'adaptation, ou par maximum de vraisemblance via un algorithme de type *Expectation-Maximisation* (Kneser et Steinbiss, 1993 ; De Mori et Federico, 1999). Selon les travaux, la distribution P_B est interpolée soit avec la seule distribution P_a (Tür et Stolcke, 2007), soit avec celles de modèles de langue appris respectivement pour un ensemble de thèmes définis *a priori* (Seymore et Rosenfeld, 1997). Outre sa simplicité, l'interpolation linéaire présente la propriété intéressante de résulter en une distribution P_A dont la qualité ne peut pas être pire que celle de la moins bonne des distributions P_i – c'est d'ailleurs en ce sens que Donald Rosenfeld écrit que l'interpolation ne peut pas faire de mal[17] (Rosenfeld, 2000). Il ne faut toutefois pas perdre de vue que cela signifie seulement que P_A ne peut pas être pire que la moins bonne des distributions entre P_B et P_a et que, si P_a est beaucoup plus mauvaise que P_B, l'interpolation linéaire peut conduire à un échec de l'adaptation. Par ailleurs, l'interpolation linéaire présente comme inconvénient de se baser sur des coefficients d'interpolation constants quelque soit h. Ceci implique qu'il n'est pas fait le meilleur usage possible des distributions interpolées. Par exemple, dans le cas d'une adaptation thématique, une probabilité $P_a[w|h]$ estimée à partir des données d'adaptation aura toujours le même poids par rapport à $P_B[w|h]$, que le n-gramme hw soit quelconque ou qu'il soit spécifique au thème en question. Il a toutefois été montré que l'interpolation linéaire pouvait être améliorée en faisant dépendre les coefficients d'interpolation de chaque historique (Weintraub *et al.*, 1996). Plus catégoriquement, certains travaux proposent même de ne considérer que les probabilités P_a pour les n-grammes ayant suffisamment été observés dans les données d'adaptation et de se replier sur les probabilités de P_B uniquement dans le cas contraire (Besling et Meier, 1995 ; Klakow, 2006), soit :

$$P_A[w|h] = \begin{cases} P_a[w|h] & \text{si } C_a(hw) \geqslant \tau, \\ Z \times P_B[w|h] & \text{sinon} \ , \end{cases} \tag{3.2}$$

où Z permet de garantir que P_A est une vraie distribution de probabilités et τ est déterminé empiriquement. Enfin, il a également été proposé d'utiliser une interpolation log-linéaire plutôt que linéaire (Gutkin, 2000) :

$$P_A[w|h] = \frac{1}{Z(h)} \times \prod_i P_i[w|h]^{\lambda_i} \ , \tag{3.3}$$

[17] « *Interpolation cannot hurt* ».

où $Z(h)$ est un coefficient de normalisation.

D'autres techniques proposent d'effectuer la combinaison au niveau des comptes de n-grammes plutôt qu'au niveau des probabilités. Ces techniques s'appuient sur une recherche par maximum *a posteriori* (MAP) où le but est de trouver la distribution qui permet de prédire le mieux possible les données d'adaptation mais qui, dans le même temps, ne s'éloigne pas trop de la distribution initiale P_B. La distribution P_A se calcule alors par l'interpolation linéaire des comptes C_B issus du corpus généraliste et ceux C_a du corpus d'adaptation (Federico, 1996), soit :

$$P_A[w|h] = \frac{C_B(hw) + \lambda \times C_a(hw)}{C_B(h) + \lambda \times C_a(h)} \ , \tag{3.4}$$

où λ est un coefficient à déterminer empiriquement. Encore une fois, il a été montré que λ pouvait être rendu dépendant de chaque historique h de manière à prendre en compte des caractéristiques linguistiques des n-grammes (Chen et Huang, 1999).

Malgré tout, ces approches visant à combiner directement des probabilités ou des comptes provenant de différentes sources offrent peu de possibilités pour explicitement contrôler les modifications attendues dans le modèle de langue initial. D'autres méthodes s'inspirent plutôt de la philosophie des modèles exponentiels (section 2.1.2.4) pour proposer un cadre d'adaptation plus souple dit par minimum d'information discriminante.

3.2.2 Adaptation par minimum d'information discriminante

L'objectif de l'adaptation par minimum d'information discriminante (MDI) est de trouver un modèle de langue dont la distribution de probabilités P_A satisfait un ensemble de contraintes et qui possède une entropie relative, ou divergence de Kullback-Leibler[18], minimale avec la distribution P_B du modèle initial (Della Pietra *et al.*, 1992). L'adaptation MDI présente alors l'intérêt de pouvoir contraindre le modèle adapté à respecter des caractéristiques statistiques explicitement définies pour la tâche d'adaptation, ces caractéristiques étant généralement observées à partir d'un corpus d'adaptation.

Ce type d'adaptation reprend directement le cadre théorique de l'apprentissage par maximum d'entropie, dans le sens où cet apprentissage peut être vu comme une adaptation MDI à partir d'une distribution initiale uniforme. Les probabilités adaptées *via* MDI s'expriment sous la forme générale :

$$P_A[w|h] = \frac{P_B[w|h] \times \alpha(hw)}{Z(h)} \ , \tag{3.5}$$

où $Z(h)$ est un coefficient de normalisation et $\alpha(hw)$ est un facteur de mise à l'échelle dépendant de chaque n-gramme hw. Ce facteur est très généralement approximé par le rapport des probabilités *jointes* données à un n-gramme hw à partir des données d'adaptation (P_a) et par le modèle généraliste à adapter (P_B) :

$$\alpha(hw) = \frac{P_a[h,w]}{P_B[h,w]} \ . \tag{3.6}$$

En pratique, $P_a[h,w]$ ne peut souvent pas être estimée de manière fiable pour tous les n-grammes car les données d'apprentissage sont généralement en quantité trop faible. Ainsi,

[18] En anglais, on emploie également le terme *discriminant information* pour désigner cette notion, d'où le nom de l'adaptation.

la quasi majorité des travaux se limitent à l'estimation de probabilités unigrammes (Kneser *et al.*, 1997 ; Federico, 1999 ; Klakow, 2006). Cette simplification conduit à mettre à l'échelle chaque probabilité $P_B(w|h)$ par un facteur ne dépendant que de w :

$$\forall h, \alpha(hw) = \alpha(w) = \frac{P_a[w]}{P_B[w]} \ . \tag{3.7}$$

Ce cas particulier d'adaptation MDI est connu sous le nom d'*unigram rescaling*. Il s'agit de la stratégie d'adaptation qui présente globalement les meilleures performances, notamment par rapport aux stratégies à base d'interpolation linéaire (Chen *et al.*, 2004). Plus précisément, sur la base d'un modèle de langue initial conduisant à un WER d'environ 20 %, les modèles adaptés d'après une technique d'*unigram rescaling* aboutissent approximativement à une diminution absolue du WER de $0,5$ à $0,8$ point pour une diminution relative moyenne de la perplexité de 15 %.

À notre sens, l'*unigram rescaling* présente toutefois l'inconvénient majeur de ne pas tirer parti du pouvoir qu'offrent les contraintes d'exprimer les caractéristiques linguistiques de la tâche d'adaptation considérée. Certaines approches tentent alors d'affiner la mise à l'échelle des probabilités n-grammes. Par exemple, certains travaux cherchent à ne pas systématiquement se ramener à une probabilité unigramme lorsque la probabilité d'un n-gramme estimée à partir des données d'adaptation est jugée suffisamment fiable d'après une mesure de significativité statistique (Chueh et Chien, 2008), cette technique conduisant à des améliorations sur le WER et la perplexité légèrement supérieures à celles obtenues par un technique d'*unigram rescaling*. D'autres travaux sur l'adaptation thématique proposent quant à eux de ne pas estimer les probabilités unigrammes P_a directement à partir des données d'adaptation mais suggèrent de construire cette distribution par des méthodes d'analyse sémantique latente (Gildea et Hofmann, 1999 ; Federico, 2002 ; Tam et Schultz, 2006 ; Tam et Schultz, 2007). Le principe de ces dernières méthodes est de projeter des données représentatives du thème (un corpus d'adaptation ou plus directement une transcription automatique) dans un espace sémantique dans lequel ces données, modélisées comme des sacs de mots, sont généralisées sous la forme d'une combinaison de variables latentes assimilables à des concepts (Deerwester *et al.*, 1990 ; Dumais, 1994). Cette représentation permet de calculer la probabilité de chaque mot du vocabulaire du système d'appartenir au thème considéré. Cette approche s'avère ainsi intéressante car elle permet de mettre en avant les mots importants pour le thème d'un document donné en généralisant le contenu lexical des données d'adaptation. Toutefois, elle n'a, à notre connaissance, jamais été comparée à l'utilisation de probabilités directement estimées à partir d'un corpus et, de plus, elle se limite encore une fois à la considération de mots seuls et non pas de séquences de mots et ne permet donc pas de prendre complètement en compte les particularités linguistiques d'un thème (expressions typiques, emplois particuliers de certains mots...).

3.2.3 Modèles de langue adaptatifs

Certaines approches proposent de ne pas adapter le modèle de langue avant chaque nouvelle transcription d'un document mais d'effectuer son adaptation dynamiquement, pendant le décodage, au fur et à mesure du traitement des groupes de souffles successifs. Pour cela, ces travaux s'appuient sur des modèles de langue dits adaptatifs au sein desquels la probabilité conditionnelle d'un mot se calcule sachant son historique n-gramme classique et un historique beaucoup plus étendu représenté sous la forme d'un sac de mots, second historique que nous distinguons du premier par le terme *cache*. Ce cache a l'intérêt de donner à

chaque instant une description du contenu lexical d'un document en cours de transcription et permet la prise en compte de relations de cause à effet entre des mots physiquement distants au sein d'un même discours.

Historiquement, ces caches ont été utilisés pour augmenter les chances d'apparition des mots ayant déjà été prononcés, partant du constat que des mots utilisés une première fois ont tendance à l'être à nouveau par la suite (Kuhn et De Mori, 1990). Certains modèles dits *à amorces* (ou *trigger models*) proposent de généraliser cette stratégie en intégrant des relations sémantiques entre les mots lors de l'apprentissage du modèle. Leur but est de favoriser la probabilité d'apparition d'un certain mot w si certains mots apparaissant habituellement dans le même contexte que lui sont présents dans le cache (Rosenfeld, 1996). Ces mots dits *déclencheurs* peuvent typiquement être appris automatiquement par des techniques de recherche d'affinités entre mots (Grefenstette, 1994). Un exemple souvent cité pour étayer cette approche est de souligner que si le mot « prêt » a été évoqué dans le discours, alors l'évocation par la suite du mot « banque » devient très probable. Depuis, ce principe a été généralisé par l'emploi déjà évoqué de techniques d'analyse sémantique latente en utilisant le cache comme point de départ pour généraliser le thème du document et déterminer des probabilités d'apparition de mots liés à ce thème (Bellegarda, 2000). Malgré des améliorations notables en terme de perplexité et de WER, ces diverses approches sont toutefois difficiles à concilier avec un décodage en faisceau efficace en raison de la mise à jour constante des probabilités conditionnelles du modèle de langue et du nombre considérable de caches actifs que le décodeur doit considérer à chaque instant.

Comme nous l'avons montré, beaucoup de techniques permettent l'adaptation d'un modèle de langue. Parmi celles-ci, le cadre des modèles exponentiels est très intéressant car il offre la possibilité de contraindre explicitement les probabilités des distributions adaptées. Toutefois, il ressort que très peu de méthodes profitent de cette possibilité pour intégrer des caractéristiques spécifiques à une tâche donnée. Plus particulièrement, peu d'efforts sont faits pour déterminer explicitement les caractéristiques lexicales et linguistiques d'un thème. En sus de la réestimation du modèle de langue, ces caractéristiques obligent également à réfléchir à des mécanismes d'adaptation du vocabulaire.

3.3 Adaptation du vocabulaire

Bien que la taille du vocabulaire d'un système de transcription n'ait en théorie pas de limite, les systèmes actuels s'en tiennent pour des raisons pratiques à des vocabulaires de plusieurs dizaines de milliers de mots. En effet, outre la difficulté d'estimer la probabilité de mots peu fréquents et le fait que de nouveaux mots apparaissent tous les jours, augmenter la taille d'un vocabulaire entraîne une plus grande confusion acoustique entre mots (Rosenfeld, 1995) ainsi qu'une augmentation de la perplexité du modèle de langue (Auzanne *et al.*, 2000), aboutissant finalement à des taux de reconnaissance moindres. Par ailleurs, disposer de vocabulaires toujours plus grands conduit rapidement à une explosion des besoins en mémoire et temps de calcul.

Face à cette limitation, le choix d'un vocabulaire qui recense un maximum des mots prononcés dans les documents à transcrire est un facteur essentiel à l'obtention de transcriptions de qualité. Ce critère de couverture lexicale dépend cependant bien entendu de la tâche considérée, un vocabulaire ayant une bonne couverture pour transcrire des émissions de sports aura probablement une couverture beaucoup plus faible lorsqu'il s'agira de trans-

crire des débats politiques. En pratique, pour la construction d'un système généraliste, un compromis est réalisé en maximisant la couverture lexicale du vocabulaire sur un ensemble de textes aux thématiques variées.

Il n'en reste pas moins que la couverture d'un vocabulaire n'est jamais complète. Il existera toujours des mots que le système ne connaîtra pas. Ces mots dit *hors vocabulaire* sont alors fatalement mal transcrits et sont généralement remplacés en sortie par une séquence de mots plus courts dont la prononciation globale s'approche de celle du mot hors vocabulaire. Plus précisément, la communauté s'accorde sur ce point pour dire que chaque occurrence d'un mot hors vocabulaire est responsable d'environ $1,5$ à 2 erreurs de reconnaissance (Rosenfeld, 1995 ; Hetherington, 1995 ; Bazzi, 2002). Ces erreurs sont d'ailleurs d'autant plus gênantes que les mots hors vocabulaire sont fréquemment des mots discriminants pour la description d'un document (entités nommées, termes techniques...). Pour la tâche de transcription de brèves d'actualités dans des langues telles que le français ou l'anglais, beaucoup de travaux s'accordent plus particulièrement sur le fait que les entités nommées, et notamment les noms de personnes, sont la catégorie majoritairement touchée par le problème des mots hors vocabulaire (Allauzen, 2003 ; Palmer et Ostendorf, 2005). Il est toutefois probable que ce constat soit moins net pour d'autres types de documents, notamment ceux où une thématique marquée est développée en détails (présentations scientifiques, émissions d'analyse...).

La question des mots hors vocabulaire est connue dans de multiples domaines. Elle a notamment été étudiée dans le cadre de la reconnaissance automatique de parole à grand vocabulaire (Geutner *et al.*, 1998a ; Allauzen et Gauvain, 2005a ; Martins *et al.*, 2006 ; Boulianne *et al.*, 2006) comme à vocabulaire restreint pour des tâches précises (Chung *et al.*, 2004 ; Scharenborg *et al.*, 2007), dans le cadre de recherche d'information dans des documents audio (Auzanne *et al.*, 2000 ; Seide *et al.*, 2004), pour des tâches de classification de textes (Batista *et al.*, 2008) ou encore de reconnaissance optique de caractères (Bazzi *et al.*, 1999). Malgré la variété des angles d'attaque adoptés, le problème des mots hors vocabulaire s'exprime à travers trois tâches distinctes. Certains travaux cherchent à détecter les portions d'une transcription susceptibles de correspondre à des mots hors vocabulaire mal transcrits (Palmer *et al.*, 2000 ; Ketabdar *et al.*, 2007 ; Burget *et al.*, 2008). D'autres travaux visent à dresser des listes de mots-candidats à intégrer dans le vocabulaire d'un système, le but étant de minimiser le taux de mots hors vocabulaire du nouveau vocabulaire pour la tâche visée tout en contrôlant sa taille. Enfin, d'autres encore s'interrogent sur la manière d'intégrer de nouveaux mots dans un système pré-existant, l'objectif étant alors de retranscrire au mieux l'ajout de nouveaux mots dans le vocabulaire en terme d'amélioration du WER. Dans le cadre de notre problématique d'adaptation thématique, nous nous intéressons uniquement à ces deux dernières tâches et laissons de côté le problème de la détection de mots mal transcrits.

3.3.1 Recherche de mots-candidats

La recherche de mots-candidats consiste à dresser une liste de mots actuellement absents du vocabulaire du système qui seraient susceptibles d'être prononcés dans un document à transcrire. En pratique, cette recherche consiste à sélectionner les mots jugés les plus pertinents pour la tâche considérée parmi l'ensemble des mots hors vocabulaire présents dans des données d'adaptation (section 3.1).

Selon l'origine de ces ressources, cette sélection est d'ailleurs plus ou moins cruciale.

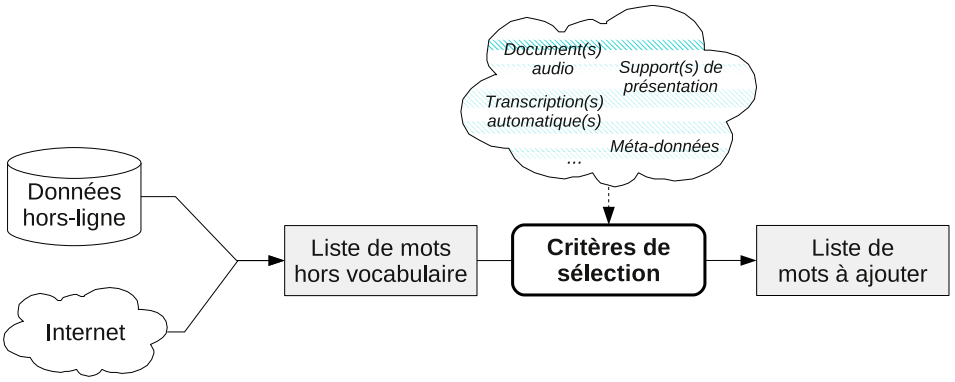


FIGURE 3.2 – Schéma de la recherche des mots à ajouter au vocabulaire.

Dans (Schwarm *et al.*, 2004), les auteurs décident par exemple d'exclure tout texte provenant d'Internet pour l'adaptation d'un vocabulaire car ils estiment que la qualité orthographique moyenne de ces textes conduirait à ajouter beaucoup de mots erronés dans le système. À l'inverse, le travail de (Kemp et Waibel, 1998), qui se base sur des articles de journaux électroniques à l'orthographe plus fiable, n'utilise aucun critère de sélection et inclut l'ensemble des mots hors vocabulaire rencontrés dans ces textes. Ni l'une ni l'autre de ces solutions n'est cependant réaliste : alors que la première position est relativement excessive, la seconde a l'inconvénient de sélectionner beaucoup trop de mots. Une solution intermédiaire consiste alors à recourir à un ou plusieurs critères de sélection pour filtrer l'ensemble des mots-candidats selon un schéma décrit par la figure 3.2. Nous en donnons ici une liste parmi la multitude de ceux relevés dans la littérature.

Critère phonétique. En considérant l'adaptation du vocabulaire d'un système comme la recherche de mots qui pourraient être prononcés, le critère de sélection le plus immédiat est de rechercher phonétiquement la présence éventuelle de certains mots-candidats dans le ou les documents à transcrire. En générant des phonétisations de chaque mot-candidat, (Geutner *et al.*, 1998b ; Palmer et Ostendorf, 2005) proposent d'aligner chacune de ces phonétisations avec des séquences de phonèmes correspondant à des mots ou séquences de mots d'une transcription automatique. Remarquons que ce critère de sélection s'apparente alors à une tâche de détection de mots-clés (*keyword spotting*) (Cardillo *et al.*, 2002) bien que, à notre connaissance, aucun travail de ce dernier domaine n'ait visé le problème de l'adaptation du vocabulaire d'un système de reconnaissance automatique de la parole.

Contexte linguistique. D'autres travaux se fondent sur le postulat selon lequel l'apparition de mots donnés dépend de celle d'autres mots ou séquences de mots « déclencheurs ». Cette contrainte s'exprime à deux niveaux : au niveau syntaxique puisqu'il existe des règles plus ou moins figées pour la construction des syntagmes d'une langue ; et au niveau paradigmatique puisque, dans un discours cohérent, le choix des mots fait par un locuteur dépend de multiples facteurs contextuels tels le thème abordé, le style oral ou encore l'époque – facteurs qui permettent implicitement à l'humain de lever des ambiguïtés d'interprétation. Dans (Oger *et al.*, 2008), ne sont retenus que les mots-candidats qui apparaissent dans des séquences à trous que l'on retrouve dans une transcription automatique d'un document audio. Dans (Ohtsuki *et al.*, 2005 ; Tam et Schultz, 2006), une technique d'analyse sémantique latente permet de calculer des vecteurs de concepts pour chaque mot-candidat.

Un mot-candidat est alors conservé si son vecteur est suffisamment proche du vecteur de concepts moyen calculés à partir des mots d'une transcription automatique.

Critère fréquentiel. Le vocabulaire de départ d'un système de reconnaissance étant principalement construit d'après un critère de fréquence d'apparition, certains travaux reprennent directement ce principe en ne sélectionnant que les quelques mots-candidats les plus fréquents dans les données d'adaptation (Auzanne *et al.*, 2000 ; Bertoldi et Federico, 2001). Plus élégamment, dans (Allauzen et Gauvain, 2005b), les comptes des corpora d'adaptation et du corpus d'apprentissage sont interpolés avant de déterminer un nouveau vocabulaire complet. Cette interpolation se fait sous la contrainte que le vecteur interpolé doit être le plus proche possible du vecteur mesuré sur un corpus cible représentatif de la tâche à modéliser.

Critère morphologique. Pour des mots peu fréquents partageant une même racine morphologique, il n'est pas rare que seules certaines flexions soient présentes dans un corpus. Par exemple, rares sont les verbes dont toutes les formes conjuguées sont répertoriées dans le vocabulaire[19]. Différents travaux proposent alors de sélectionner les mots-candidats dont une forme fléchie ou morphologiquement proche est présente dans une transcription automatique (Geutner *et al.*, 1998a ; Arısoy et Saraçlar, 2009). Pour aller plus loin, (Martins *et al.*, 2006) propose de générer automatiquement des variantes de verbes. Dans l'idéal, cette approche devrait être étendue à toutes les catégories et aux transformations morphologiques dérivationnelles (préfixation, suffixation...). Cependant, cette généralisation est une tâche complexe, qui plus est pour les langues comme le français qui présentent un grand nombre d'exceptions. Par ailleurs, on pourrait craindre de ce genre de techniques qu'il génère trop de mots erronés et que le nombre de mots à ajouter augmente trop vite.

Critère thématique. Dans l'ensemble, on note que peu de travaux prennent en compte des connaissances liées au thème d'un document. À notre connaissance, seul le travail présenté dans (Marin *et al.*, 2009) cherche à évaluer l'importance thématique de mots-candidats par le calcul de multiples variantes du scores *tf-idf*, scores visant à évaluer le pouvoir discriminant d'un mot pour distinguer le texte auquel il appartient par rapport à un ensemble d'autres textes traitant de sujets différents. Cette unique tentative s'avère toutefois sans succès pour sélectionner efficacement de nouvaux mots. Malgré tout, il nous semble que des efforts devraient être poursuivis dans cette voie, étant donné que, au sein d'un thème, l'emploi de mots techniques est relativement fréquent.

La sélection de mots hors vocabulaire à ajouter à un vocabulaire est une étape primordiale dans l'adaptation d'un vocabulaire. En pratique, l'efficacité des différentes techniques proposées pour ce problème dépend largement de la langue considérée, le taux de mots hors vocabulaire d'un vocabulaire initial atteignant facilement 10 % pour des langues très flexionnelles ou agglutinantes alors qu'il est plus généralement autour de 1 ou 2 % pour des langues proches du français. Dans ce dernier cas, les meilleures techniques d'adaptation arrivent généralement à diviser ce taux par deux. Dans le cadre d'un système de reconnaissance automatique de la parole, le travail ne s'arrête cependant pas là. Il reste encore à

[19] Ce constat est encore plus flagrant pour des langues agglutinantes ou fortement flexionnelles, où le nombre de variantes d'une même racine est très grand.

intégrer au mieux les mots sélectionnés dans le modèle de langue afin que leur présence se fasse ressentir en terme de gain sur le taux de reconnaissance.

3.3.2 Intégration de nouveaux mots dans le système

Outre le référencement de chaînes graphiques, l'intégration de nouveaux mots dans un système de reconnaissance automatique de la parole passe par deux étapes : la génération de prononciation pour chacun d'entre eux et leur intégration dans le modèle de langue du système. Comme nous l'avons déjà vu, la première étape peut être effectuée par le recours à des outils de phonétisation automatique, bien que ces derniers ne soient pas totalement fiables. L'étape de mise à jour du modèle de langue reste plus ouverte quant à la méthode à adopter. Deux techniques sont toutefois récurrentes dans la littérature : le réapprentissage du modèle de langue ou l'utilisation d'un modèle initial comprenant une ou plusieurs classes ouvertes à mettre à jour. Par ailleurs, d'autres approches proposent de ne pas modifier le modèle de langue mais d'intégrer directement les nouveaux mots au niveau des sorties d'un décodage. Nous donnons un aperçu de l'ensemble de ces techniques.

3.3.2.1 Réapprentissage d'un modèle de langue fermé

Disposant d'un modèle n-gramme initial, la prise en compte d'un nouveau vocabulaire impose la réestimation complète ou partielle du modèle de langue du système. Dans ce cas, l'intégration de nouveaux mots dans le système rejoint la problématique de l'adaptation du modèle de langue. Ainsi, certains travaux choisissent de mélanger des comptes de n-grammes issus du corpus d'apprentissage avec ceux du corpus d'où ont été tirés les nouveaux mots du vocabulaire, corpus alors vu comme un corpus d'adaptation (Kemp et Waibel, 1998 ; Allauzen et Gauvain, 2005a ; Yamazaki *et al.*, 2007). D'autres travaux procèdent plus modestement à l'apprentissage d'un « petit » modèle de langue sur le corpus d'adaptation avant de l'interpoler avec le modèle de langue initial (Geutner *et al.*, 1998a ; Auzanne *et al.*, 2000 ; Ogata *et al.*, 2007). Pour ne pas pénaliser les n-grammes comportant de nouveaux mots de leur absence dans le modèle initial, le choix des coefficients d'interpolation peut se fonder sur des statistiques issues d'un texte caractéristique de la tâche, par exemple une transcription automatique fournie par le système initial (Martins *et al.*, 2006).

Ces approches basées sur l'apprentissage d'un modèle de langue ont l'intérêt d'intégrer les nouveaux mots du vocabulaire dans des contextes fiables car observés dans un corpus spécialisé. Ce point est toutefois à pondérer par le fait que les corpora d'adaptation sont généralement de faible taille et qu'ils ne couvrent donc qu'un relativement faible nombre d'historiques possibles pour les nouveaux mots, d'autant plus que ces mots sont rares, y compris parfois au sein des données d'adaptation.

3.3.2.2 Modèle de langue à vocabulaire ouvert

Un modèle de langue à vocabulaire ouvert est un modèle appris sur un vocabulaire intégrant une ou plusieurs classes symbolisant l'ensemble des mots absents d'un vocabulaire initial. Ces classes sont considérées comme des membres à part entière du vocabulaire et disposent par conséquent de probabilités d'enchaînement avec les autres mots du modèle de langue. Ces modèles ont l'intérêt de permettre l'introduction de nouveaux mots dans le vocabulaire sans avoir à toucher au modèle de langue.

Dans le cas le plus simple, une seule classe représentant l'ensemble des mots absents du vocabulaire est considérée (Bertoldi et Federico, 2001 ; Ohtsuki *et al.*, 2005). Elle est généralement symbolisée par l'unité lexicale <unk>. L'ajout d'un nouveau mot dans le vocabulaire se fait alors par l'assignation de ce mot à la classe <unk> dont il apparaît alors pour le système comme une variante de prononciation. Le calcul des probabilités conditionnelles faisant intervenir ce nouveau mot w_{oov} s'apparente alors au calcul d'une probabilité dans un modèle n-classe :

$$P[w_{oov}|h] = P[\text{<unk>}|h] \times P(w_{oov}|\text{<unk>}) \ , \tag{3.8}$$

où h est un historique. En pratique, le terme $P(w_{oov}|\text{<unk>})$ permet d'éviter une surestimation des probabilités contenant la classe <unk> par rapport à celle basée sur les autres mots du vocabulaire. Ce terme est toutefois difficile à estimer car le nombre de mots appartenant potiellement à la classe <unk> est non borné. Une probabilité constante, déterminée empiriquement, est alors souvent attribuée à tous les membres de cette classe. L'utilisation d'une seule classe tend toutefois à être proscrite car cela revient à considérer que tous les mots hors vocabulaire sont syntaxiquement interchangeables, ce qui est clairement faux. Pour pallier ce problème, une majorité de travaux considèrent plutôt un ensemble de classes dont chacune correspond à une catégorie morphosyntaxique (Allauzen et Gauvain, 2005b ; Oger *et al.*, 2008) voire à un rôle sémantique déterminé (Chung *et al.*, 2004 ; Scharenborg *et al.*, 2007).

L'utilisation d'un modèle de langue à vocabulaire ouvert présente donc l'intérêt de ne pas nécessiter de réapprentissage tout en permettant de spécialiser en partie l'intégration de nouveaux mots *via* le recours à différentes classes. Toutefois, le choix d'un découpage en classes ne résulte en pratique que de l'étude d'un trait unique des mots (catégorie morphosyntaxique, genre, rôle sémantique...), ce qui revient donc malgré tout à confondre au sein de mêmes classes des mots qui se distinguent pourtant par des emplois dans des contextes différents. Par ailleurs, en tant que modèle hybride entre n-grammes et n-classes, les modèles à vocabulaire ouvert sont plus difficile à adapter linguistiquement. Il est notamment difficile d'appliquer directement les méthodes traditionnelles d'adaptation vue à la section 3.2.

3.3.2.3 Autres approches

Si les deux grandes tendances présentées plus haut nécessitent un second décodage complet pour pouvoir prendre en compte les nouveaux mots d'un vocabulaire, d'autres travaux optent pour des traitements *a posteriori*. Dans (Palmer et Ostendorf, 2005), les auteurs tentent de corriger des erreurs présumées dans la meilleure hypothèse rendue par un système initial. Pour chaque erreur supposée, un mot de remplacement est sélectionné d'après un critère phonétique parmi un ensemble de mots-candidats. Ce raisonnement est étendu à des graphes de mots dans (Arısoy et Saraçlar, 2009). Pour un graphe donné, chaque arc porteur d'un mot est mis en parallèle avec un ensemble de nouveaux arcs dont les mots répondent à un critère de similarité morphologique ou phonétique avec le mot comparé. Cependant, cette dernière méthode présente l'inconvénient majeur de devoir attribuer des scores linguistiques aux nouveaux arcs. Dans le travail évoqué, les auteurs décident de recourir à un collossal modèle de langue basé sur 670 000 mots. Nous pensons que cette démarche n'est pas idéale car l'apprentissage d'un tel modèle est vraisemblablement très coûteux en temps et en données textuelles nécessaires et, par ailleurs, il n'est clairement

pas exclu que le modèle de langue et le vocabulaire de 670 000 mots ne nécessitent pas, eux aussi, une adaptation à un moment donné.

Finalement, de nombreuses techniques sont proposées pour l'adaptation d'un vocabulaire. Alors que toutes conduisent à des améliorations des taux de reconnaissance d'un système, il reste néanmoins difficile de comparer précisément leurs propensions respectives à traduire l'ajout d'un nouveau mot en une diminution du WER du fait de la grande variabilité des cadres expérimentaux rapportés dans la littérature. Il se dégage pourtant des différents travaux que la stratégie fondée sur un vocabulaire ouvert semble la plus utilisée actuellement. Ceci s'explique probablement par son élégant mécanisme de repli lexical qui évite toute réestimation du modèle de langue tout en produisant des gains significatifs sur le WER. Dans l'ensemble, il ressort malgré tout que peu de techniques cherchent à intégrer des connaissances explicites sur les thèmes.

3.4 Bilan

Comme l'illustre cet état de l'art, il existe une profusion de travaux s'intéressant à l'adaptation des composants linguistiques d'un système de reconnaissance automatique de la parole à une tâche donnée. Ces travaux proposent des solutions pour la récupération de données d'adaptation, l'adaptation de modèles de langue et la mise à jour du vocabulaire. Bien que ces diverses études rapportent dans l'ensemble des améliorations des taux de reconnaissance et des diminutions de la perplexité des modèles de langue adaptés, nous leur adressons deux principaux reproches. D'une part, ces solutions présentent généralement des choix techniques intéressants mais intègrent peu de connaissances linguistiques liées à la tâche d'adaptation considérée. Dans le cas de l'adaptation thématique, seuls des travaux faisant appel à l'analyse sémantique latente semblent tenter de dépasser les simples données brutes des corpora d'adaptation. À notre sens, ces techniques sont toutefois limitées car elles ne considèrent que des mots isolés, et non des séquences de mots. De plus, elles s'appuient sur un cadre statistique difficilement interprétable et manipulable, ce qui s'écarte des impératifs que nous sommes fixés, à savoir disposer d'éléments explicites de sémantique. D'autre part, peu de travaux s'intéressent à l'intégralité de la chaîne d'adaptation. La faisabilité d'une telle approche dans un cadre complètement non supervisé reste pourtant une question importante.

Ces différents constats remettent en lumière l'objectif de cette thèse. Nous cherchons à proposer un schéma d'adaptation thématique complètement non supervisé au sein duquel les caractéristiques lexicales et linguistiques d'un thème sont clairement mises en évidence et exploitées. Le chapitre suivant donne une vue d'ensemble de l'approche que nous avons retenue pour mener à bien cet objectif et présente le cadre expérimental sur lequel nous nous appuyons pour juger de la qualité de nos propositions.

Chapitre 4

Approche retenue et cadre expérimental

Ce chapitre présente la démarche originale d'adaptation thématique non supervisée des composants linguistiques d'un système de transcription que nous proposons et développons dans le reste de ce manuscrit. Pour rappel, nous nous sommes fixé un cadre de travail précis pour la mise en œuvre de cette démarche. Tout d'abord, nous cherchons à construire un processus d'adaptation entièrement non supervisé. Ceci implique que nous ne souhaitons absolument pas faire d'hypothèse quant aux thèmes des documents à transcrire que le système pourrait rencontrer. Plus généralement, nous ne voulons utiliser aucune ressource ou connaissance *a priori* à moins que celle-ci puisse être générée automatiquement. Par ailleurs, nous souhaitons que la chaîne de traitements sous-jacente à notre processus d'adaptation soit complète et cohérente, c'est-à-dire que nous ne cherchons pas à proposer un assemblage de méthodes indépendantes mais posons un cadre commun à l'ensemble des traitements. En réponse à ces contraintes, nous visons alors l'intégration de techniques de traitement automatique des langues au sein de notre processus. Plus particulièrement, nous souhaitons pouvoir caractériser explicitement les spécificités linguistiques de chaque thème rencontré à travers des éléments directement interprétables, comme par exemple des mots ou des séquences de mots. Nous laissons ainsi de côté les techniques basées sur l'analyse sémantique latente car celles-ci représentent des textes comme des combinaisons de concepts parfois abscons. Enfin, nous nous imposons de rester dans le cadre des modèles n-grammes car ces derniers sont les modèles les plus performants actuellement et sont en cela ceux qu'utilise la plupart des systèmes actuels de reconnaissance automatique de la parole à grand vocabulaire.

Ces objectifs rappelés, nous décidons dans ce court chapitre de donner une vue globale du processus d'adaptation que nous avons développé afin de faciliter la compréhension des chapitres 5, 6 et 7 qui en détaillent respectivement chacune des étapes-clés. À cette fin, la section 4.1 expose la chaîne de traitements de notre processus d'adaptation alors que la section 4.2 présente le cadre expérimental commun à l'ensemble de nos travaux, cette présentation comprenant la description des documents audio que nous utilisons ainsi que celle de nos méthodes d'évaluation.

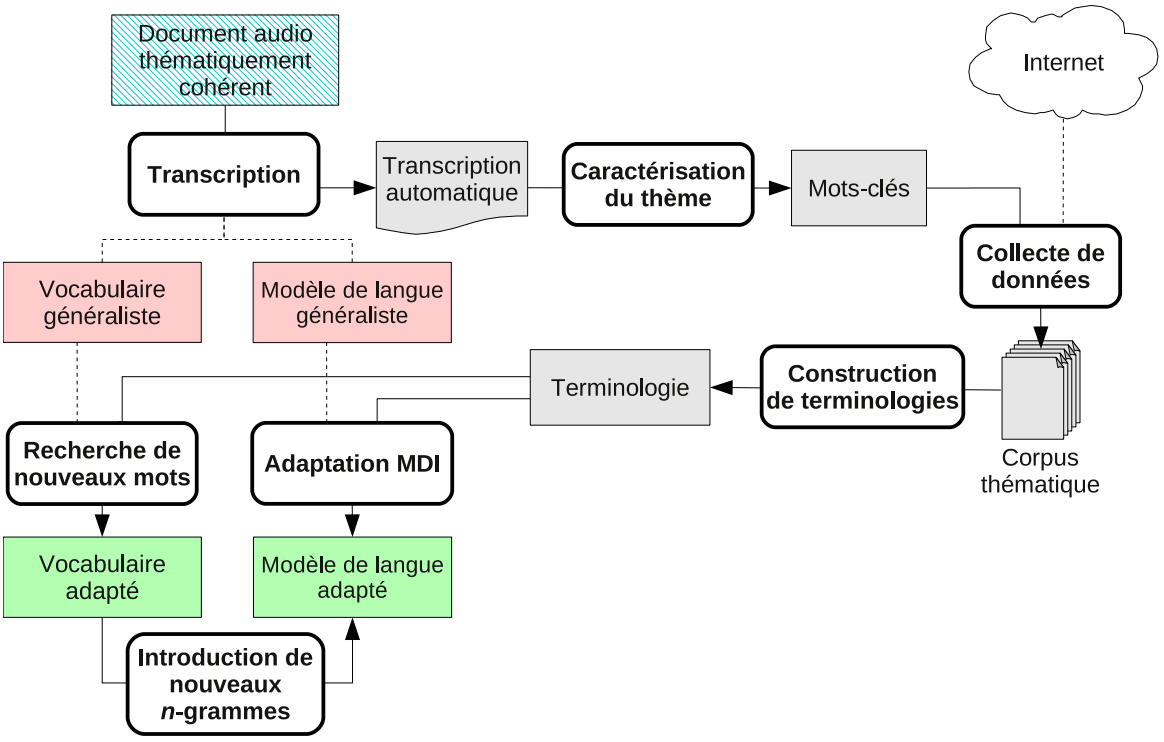


FIGURE 4.1 – Vue détaillée de notre processus d'adaptation thématique des composants linguistiques d'un système de reconnaissance automatique de la parole.

4.1 Approche retenue

Le but de la procédure d'adaptation thématique non supervisée que nous proposons est d'adapter les modèle de langue et vocabulaire généralistes d'un système de reconnaissance initial pour chaque nouveau document audio à transcrire en vue d'en fournir la meilleure transcription automatique possible. Cette adaptation se fait sous l'hypothèse que chacun de ces documents est thématiquement cohérent. Il pourrait, par exemple, s'agir de traiter les documents venant de la base d'un site de partage de vidéos, de *catch-up TV*[20] ou étant issus d'une segmentation automatique d'un flux multimédia (Naturel *et al.*, 2006 ; Guinaudeau *et al.*, 2010).

Pour chaque document audio, nous décomposons le travail en trois étapes. Tout d'abord, nous construisons un corpus textuel abordant le même thème que le document. Ensuite, nous réestimons les probabilités du modèle de langue généraliste à partir des informations fournies par le corpus thématique. Enfin, nous adaptons le vocabulaire du système en recherchant à partir du corpus thématique les mots supposés être en rapport avec le thème du document. Tel que le présente en détails la figure 4.1, nous proposons une chaîne complète, cohérente et entièrement non supervisée pour instancier ces différentes étapes. Tout d'abord, nous nous appuyons sur une première transcription générée par le système généraliste pour extraire automatiquement des mots-clés qui permettent de caractériser le thème du document audio. Ces mots-clés servent alors de support à la recherche de textes

[20] Ou « télévision de rattrapage ». Il s'agit des portails que mettent généralement en place les chaînes de télévision pour permettre visionner ou revisionner des émissions déjà diffusées sur leur antenne.

thématiquement en rapport avec le document par le biais de requêtes soumises à un moteur de recherche sur Internet. Le corpus constitué de pages Web qui en découle permet d'avoir une vision plus large des usages de la langue pour le thème du document traité. Pour mettre en évidence ces spécificités linguistiques, nous construisons ensuite automatiquement une terminologie, c'est-à-dire un ensemble de mots et séquences de mots qui portent des notions du thème du corpus. Dans un premier temps, ces connaissances sur le thème nous permettent de mettre en œuvre une adaptation MDI spécialisée pour adapter le modèle de langue généraliste. Dans un second temps, la terminologie nous permet de rechercher des mots hors vocabulaire liés au thème, mots que nous ajoutons au vocabulaire et que nous intégrons dans le modèle de langue par l'ajout de nouveaux n-grammes générés automatiquement.

Cette procédure, initiée dans (Lecorvé *et al.*, 2008b), est particulièrement originale car la notion de thème y est implicitement manipulée par l'utilisation de la pondération *tf-idf*, issue du traitement automatique des langues et du domaine de la recherche d'information (Salton, 1989) et que nous avons adaptée aux spécificités de l'oral. Cet outil, qui permet d'évaluer l'importance des mots d'un texte pour discriminer ce texte par rapport à un ensemble d'autres textes traitant de thèmes variés, instaure un cadre de travail cohérent entre toutes les étapes de notre procédure d'adaptation et rend cette dernière complètement non supervisée.

4.2 Cadre expérimental

4.2.1 Données utilisées

Les documents audio utilisés pour expérimenter notre processus d'adaptation thématique sont tirés du corpus ESTER. À l'origine, ce dernier a été constitué dans le cadre d'une campagne d'évaluation dédiée à la transcription enrichie et à l'indexation de journaux radiophoniques en Français (Gravier *et al.*, 2004). Il rassemble ainsi un ensemble d'émissions d'actualités accompagnées de leur transcription de référence. Ces émissions, datant de 1998 à 2004, proviennent de diverses radios et traitent de sujets très variés. Leur genre est principalement de la parole préparée car une majorité des locuteurs sont des journalistes ayant préalablement travaillé leurs interventions. Le corpus ESTER contient toutefois quelques passages de parole spontanée, correspondant par exemple à des interviews.

De fait, le corpus ESTER dispose d'une partie dédiée à l'apprentissage des composants du système de transcription et d'une partie dédiée aux réglages des différents paramètres (facteur d'échelle du modèle de langue, pénalité d'insertion, seuil d'élagage du décodage en faisceau...). Dans notre travail, nous avons extrait 6 heures d'enregistrements de cette seconde partie afin de constituer deux bases de documents audio thématiquement cohérents, l'une faisant office d'*ensemble de développement* pour nos diverses expériences préliminaires, l'autre étant l'*ensemble de test* sur lequel nous validerons nos travaux. Concrètement, après avoir retiré les zones correspondant aux annonces des titres, publicités et autres passages non marqués thématiquement, les 6 heures d'enregistrements ont été découpées manuellement en 172 documents thématiquement indépendants – par la suite, nous référons à ces documents par le terme *segment thématique* ou simplement *segment*. Ce découpage des documents audio provient d'un travail préalable au nôtre qui s'intéressait à la tâche de segmentation thématique automatique. Comme le montre le tableau 4.1, ces segments ont alors été répartis entre nos deux ensembles à raison de 91 pour le développement et de 81

	Radio	Date/Heure	Nb. segments	Nb. mots
Ensemble de développement	France Inter	18/04/2003 07h00-08h00	31	11 356
	France Info	18/04/2003 12h00-13h00	35	10 426
	RFI	08/05/2003 14h00-15h00	25	10 140
		Total	91	31 922
Ensemble de test	France Inter	18/04/2003 08h00-09h00	20	12 067
	France Info	18/04/2003 17h00-18h00	35	9 589
	RFI	09/05/2003 14h00-15h00	26	10 317
		Total	81	31 973

TABLE 4.1 – Composition des ensembles de développement et de test.

pour le test. Ceux-ci tirent leur origine de trois radios différentes et datent de début 2003. Ils abordent des thèmes très variés relevant aussi bien de la politique que du sport ou des sciences, et sont de tailles variables, certains segments ne contenant que quelques dizaines de mots alors que les plus longs dépassent le millier de mots. Des exemples de segments sont donnés en annexe B.1.

4.2.2 Méthodes d'évaluation

L'efficacité de notre procédure d'adaptation a été évaluée en comparant les performances obtenues par notre système avant et après adaptation pour chaque segment thématique. Pour des raisons d'interprétabilité des résultats, nous avons été amené à expérimenter séparemment l'adaptation du seul modèle de langue et celle du vocabulaire, qui, elle, nécessite toutefois également de modifier le modèle de langue. Il en découle un protocole d'évaluation, présenté par la figure 4.2, qui s'appuie sur la mesure du WER ainsi que, dans certains cas, celle de la perplexité. Après avoir appliqué ce même protocole à chacun de nos segments thématiques, nous nous ramenons à des variations moyennes de la perplexité et du WER entre le système généraliste et un système adapté, que ce soit pour notre ensemble de développement ou pour celui de test.

Lorsque seul le modèle de langue est adapté, le vocabulaire ne varie pas et il est donc possible de comparer les perplexités des modèles de langue adapté et généraliste mesurées sur la transcription de référence d'un même segment. À l'échelle d'un ensemble E de segments thématiques (l'ensemble de développement ou celui de test), la perplexité moyenne par segment est alors calculée comme :

$$\mathrm{PPL}(E) = \frac{\sum_{S \in E} \mathrm{PPL}(S)}{|E|} \ , \tag{4.1}$$

où $|E|$ est le nombre de documents dans l'ensemble E. Quant à elle, la variation moyenne de la perplexité est exprimée comme la moyenne des variations de la perplexité relevées segment par segment :

$$\Delta\mathrm{PPL}(E, M_1, M_2) = \frac{\sum_{S \in E} \mathrm{PPL}_{M_2}(S) - \mathrm{PPL}_{M_1}(S)}{|E|} \ , \tag{4.2}$$

où M_1 et M_2 sont respectivement les modèles de langue généraliste et adapté. Cette mesure permet de donner une idée de la variation de la perplexité attendue après l'adaptation du modèle de langue généraliste pour un segment thématique quelconque[21].

[21] Ce que ne permet par le calcul de la différences entre perplexités moyennes, cette dernière ne donnant qu'une vision globale des résultats à l'échelle de l'ensemble de test ou de développement

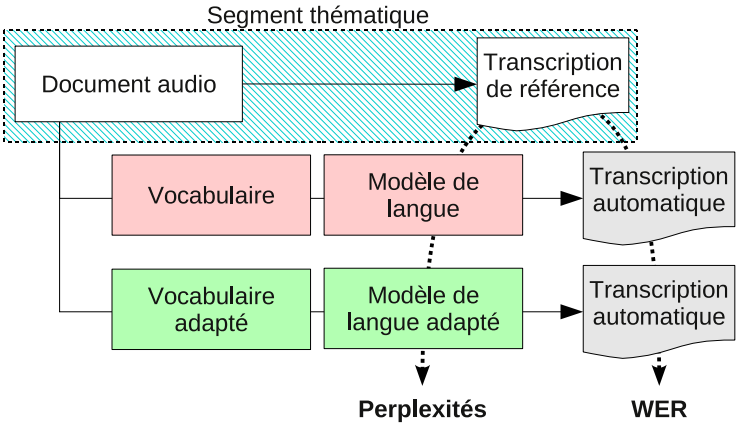


FIGURE 4.2 – Méthode d'évaluation de la procédure d'adaptation thématique.

Que le système soit adapté entièrement ou uniquement au niveau de son modèle de langue, l'objectif de notre adaptation thématique est de produire des transcriptions de la meilleure qualité possible. En cela, notre critère d'évaluation principal consiste à comparer les WER mesurés sur les transcriptions automatiques générées par le système avant et après adaptation. Nous calculons alors le WER global d'un ensemble de segments thématiques pour chacune de ces deux configurations et rapportons les variations absolues entre ces taux. Plus généralement, cette méthode d'évaluation vaut pour n'importe quel autre taux d'erreur que nous serons amené à considérer (LER, WER sur un ensemble restreint de mots...).

Ce mode d'évaluation est appliqué sur le système IRENE tout au long de nos travaux. Toutefois, il est à noter que, ce système ayant subi des modifications durant le déroulement de notre thèse, les performances rapportées d'un chapitre à l'autre ne sont pas toujours comparables dans l'absolu. Il importe principalement, tant au niveau du WER que de la perplexité, d'observer les variations de performances avant et après adaptation thématique. Dans les deux cas, plus cette variation est négative, plus l'adaptation est jugée efficace.

Afin de vérifier que ces variations ne sont pas dues au hasard, nous recourons par ailleurs à des tests de significativité statistiques. Ces tests étudient le comportement d'une variable aléatoire $Z = E_1 - E_2$ où E_1 et E_2 sont le nombre de mots mal transcrits par deux systèmes respectifs pour un même groupe de souffle (Gillick et Cox, 1989). Dans notre cas, ils évaluent la probabilité pour que l'hypothèse H_0, stipulant que les différences observées sont dues au hasard, soit rejetée en confrontant le comportement de Z observé à partir des transcriptions avec son comportement théorique prédit par une distribution donnée (par exemple, une loi normale centrée en 0). En pratique, ces tests fournissent, pour un niveau de confiance α donné, une valeur p telle que les différences observées entre les deux systèmes sont jugées statistiquement significatives si $p < \alpha$, et non significatives sinon. Dans nos expériences, nous utilisons un niveau de confiance $\alpha = 0,05$ et recourons à un test de *Student* et à un test de *Wilcoxon* (Saporta, 1990).

Maintenant que nous avons exposé notre procédure d'adaptation thématique et décrit notre cadre expérimental, les chapitres suivants s'attachent à décrire nos contributions pour chacune des grandes étapes que sont la création non supervisée de corpora thématiques, l'adaptation thématique d'un modèle de langue et d'un vocabulaire.

Chapitre 5

Création automatique d'un corpus thématique

La construction d'un corpus thématique est une étape essentielle à notre démarche d'adaptation puisque ce corpus doit constituer une ressource suffisante pour caractériser précisément les pratiques linguistiques liées au thème considéré. Comme nous l'avons souligné lors des chapitres précédents, nous souhaitons que cette étape soit complètement non supervisée dans le sens où l'ensemble des traitements doit être automatique et où aucune hypothèse ne doit être faite sur la nature des thèmes pour lesquels une adaptation doit être effectuée. Cette volonté met en avant deux besoins principaux. D'une part, les documents des corpora thématiques doivent venir d'une ressource textuelle couvrant le plus grand nombre de thèmes possible pour s'assurer de la réussite de l'adaptation thématique dans un maximum de cas et, d'autre part, il est nécessaire de savoir caractériser un thème sans recourir au jugement humain ni se baser sur une classification *a priori* en grands thèmes. Pour cela, notre approche vise à se servir d'Internet comme d'une ressource linguistique ouverte et à recourir à des techniques de traitement automatique des langues, notamment des techniques utilisées dans le domaine de la recherche d'information.

La figure 5.1 présente les détails de cette approche originale (Lecorvé *et al.*, 2008a). Pour un segment thématique donné dont une transcription automatique nous est fournie par le système de reconnaissance généraliste, nous cherchons dans un premier temps à extraire des mots-clés de manière à caractériser le thème du segment. Ces mots sont ensuite utilisés pour construire des requêtes qui sont soumises à un moteur de recherche sur Internet[22]. Les pages Web ainsi récupérées sont finalement filtrées de manière à constituer un corpus thématique en cohérence avec le segment traité. Pour évaluer la qualité et le potentiel des corpora produits, nous restreignons pour l'instant notre problème d'adaptation thématique d'un système de reconnaissance à la seule tâche d'adaptation de son modèle de langue généraliste. Pour cela, nous nous appuyons sur une stratégie classique que nous ne cherchons pas à optimiser outre mesure : pour chaque segment thématique, nous estimons un modèle de langue d'adaptation à partir du corpus thématique que nous aurons construit automatiquement avant d'interpoler linéairement ce modèle avec le modèle de langue généraliste du système. L'objectif est alors de minimiser la perplexité et le WER obtenus à partir du modèle adapté en faisant varier les différents paramètres que nous introduisons au cours

[22] Dans nos expériences, *Yahoo !* et *Bing* (anciennement *Live Search*).

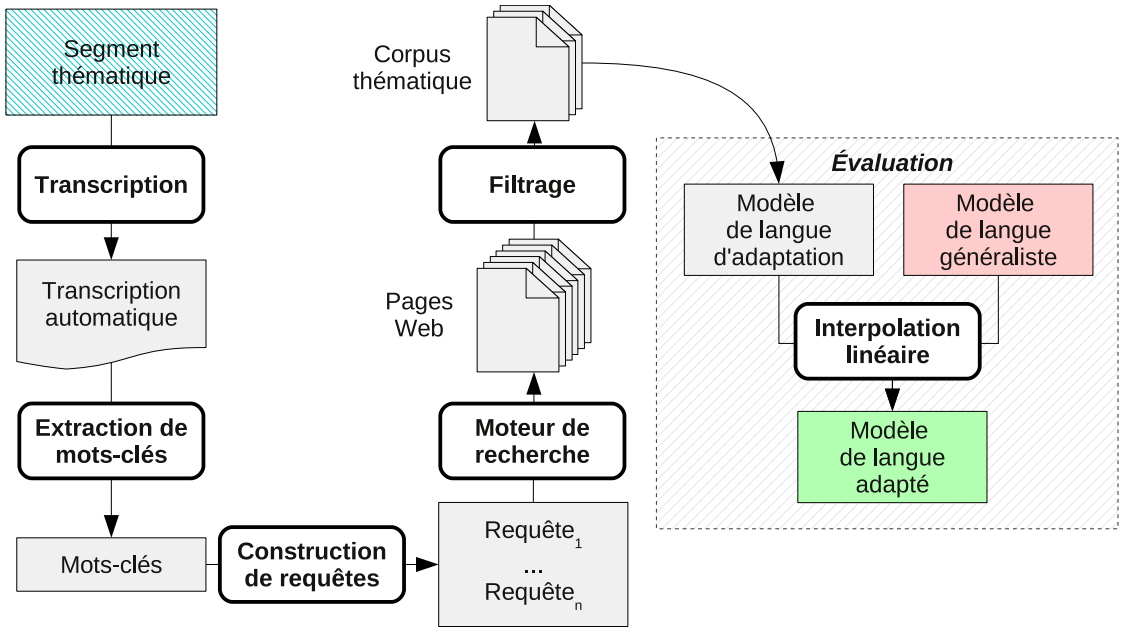


FIGURE 5.1 – Schéma de notre processus de création non supervisée d'un corpus thématique.

de ce chapitre pour notre procédure de création d'un corpus. Le calcul d'un taux d'erreur impliquant néanmoins de lancer un processus de transcription, nous privilégions le calcul de perplexités durant nos expériences préliminaires sur l'ensemble de développement et réservons le calcul de taux d'erreur à l'évaluation finale sur notre ensemble de test.

La chaîne de traitements que nous proposons pose de nombreuses questions. Tout d'abord, comment extraire des mots-clés qui caractérisent suffisamment tous les aspects d'un segment mais qui, au sein d'une requête, restent suffisamment généraux pour aboutir à un nombre conséquent de pages Web ? Comment gérer le cas d'éventuelles erreurs de transcription ? Comment assembler ensuite les différents mots-clés pour former une ou plusieurs requêtes ? Enfin, il faut également s'interroger sur la manière de sélectionner les pages pertinentes parmi toutes celles récupérées pour obtenir un corpus d'adaptation suffisamment grand et thématiquement homogène.

Ce chapitre présente nos réponses à ces différentes questions à travers la présentation détaillée des étapes d'extraction de mots-clés (section 5.1), de formulation de requêtes (section 5.2) et de filtrage thématique de pages Web (section 5.3). Ce chapitre est complété par l'étude de l'intérêt de notre procédure pour différentes tâches en section 5.4. Principalement, nous évaluons la qualité de nos corpora thématiques pour améliorer les performances de notre système de transcription *via* l'adaptation de son seul modèle de langue. Mais nous montrons également que notre chaîne de traitements peut se révéler profitable pour créer des descriptions riches de contenus multimédias. Finalement, nous donnons des pistes d'amélioration de notre méthode en section 5.5.

5.1 Extraction de mots-clés

La première étape de notre approche consiste à caractériser précisément le thème d'un segment de manière à récupérer des pages Web abordant le même thème. Pour cela, notre

idée est d'extraire des mots-clés à partir de la transcription automatique de ce segment. En nous appuyant directement sur le contenu du segment, cette technique offre l'avantage de ne nécessiter aucune hypothèse sur les thèmes pouvant être rencontrés. Nous proposons de réaliser cette tâche d'extraction de mots-clés en nous appuyant sur le critère *tf-idf* (Salton, 1989), critère bien connu dans le domaine de la recherche d'information. Cependant, ce critère a initialement été conçu pour l'indexation de textes écrits classiques et non pour la caractérisation du thème de transcriptions automatiques. Il doit donc être adapté pour prendre en compte les spécificités de ces transcriptions, à savoir la présence potentielle de mots erronés ainsi que l'absence de casse et de ponctuation, et pour permettre la caractérisation du thème d'un document. Cette section introduit tout d'abord le critère *tf-idf* original puis présente les modifications que nous y avons apportées.

5.1.1 Critère *tf-idf*

Le principe du critère *tf-idf* est d'associer à chaque mot d'un texte un score qui exprime son importance pour discriminer le texte par rapport à une collection de référence de textes. Étant donné un mot w d'une transcription t, ce score est calculé comme le produit d'une pondération locale $tf(w,t)$, qui exprime la fréquence de w dans t, et d'une pondération globale $idf(w,\Omega)$ liée au nombre de documents de la collection de référence Ω[23] qui contiennent le mot w. Parmi les diverses variantes existant pour la mise en œuvre de ces deux paramètres (Singhal, 1997), nous calculons ces fonctions de la manière suivante :

$$tf(w,t) = \frac{freq(w,t)}{\max\limits_{x\in t} freq(x,t)} \quad \text{avec} \quad freq(w,t) = \frac{|w|_t}{|t|} \tag{5.1}$$

$$\text{et} \quad idf(w,\mathcal{C}) = \log \frac{|\Omega|}{|w|_\Omega}, \tag{5.2}$$

où $|t|$ désigne le nombre de mots de t, $|\Omega|$ désigne le nombre de documents dans la collection Ω, et $|w|_t$ et $|w|_\Omega$ désignent respectivement le nombre d'occurrences de w dans t et le nombre de documents de Ω qui contiennent le mot w. Finalement, après normalisation, on obtient le score $S(w)$ compris entre 0 et 1 :

$$S(w) = \frac{tf(w,t) \times idf(w,\Omega)}{\max\limits_{x\in t} tf(x,t) \times idf(x,\Omega)}. \tag{5.3}$$

À l'exception des mots vides[24], ce score est calculé pour chaque mot de t, les mots ayant les scores les plus élevés étant alors jugés comme les plus discriminants. Ce sont ces quelques mots que nous considérons comme des mots-clés caractéristiques du thème d'un segment.

Par ailleurs, comme certaines transcriptions sont très courtes et ne contiennent que peu de répétitions de mots, les scores *tf-idf* peuvent s'avérer insatisfaisants tels quels. Pour pallier ce problème, nous appliquons une lemmatisation sur chaque transcription afin de rassembler au sein d'une même classe les mots partageant un même lemme. Ainsi, pour un lemme ℓ, nous calculons $S(\ell)$ à partir des versions lemmatisées de la transcription et

[23] En pratique, nous utilisons comme collection de référence les mêmes 800 000 articles du journal *Le Monde* datés de 1987 à 2003 qui ont contribué à l'estimation de notre modèle de langue généraliste.

[24] Les mots vides sont des mots non significatifs pour la description d'un texte, par exemple des prépositions, des pronoms, des interjections... Ces mots sont usuellement recensés au sein d'une liste qu'on appelle *stop list* et sont à opposer aux *mots pleins* ou *mots lexicaux*.

de collection de référence. Toutefois, souhaitant effectuer des requêtes dans un moteur de recherche à partir de mots réellement observés dans la transcription, nous conservons en mémoire une association entre chaque lemme et sa flexion la plus fréquente dans la transcription. Pour résumer, chaque lemme peut donc être perçu dans notre travail comme une classe d'équivalence entre mots, réprésentée par son mot le plus fréquent dans le texte.

La table 5.2 présente un exemple de 10 mots présentant le meilleur score $S(\ell)$ à partir du segment thématique de la table 5.1 évoquant le problème du port d'un voile religieux sur les photos d'identité en France. Au sein de cette liste, nous pouvons d'ores et déjà remarquer la présence de mots très caractéristiques du thème abordé. Nous retrouvons par exemple les mots voile, laïcité, musulman ou encore photo. Cependant, d'autres mots sont moins pertinents (mots en gras dans le tableau). Il s'agit majoritairement de noms propres et d'erreurs de transcriptions introduites par le système. Bien que ces mots-clés puissent être intéressants pour certaines applications, ils s'avèrent problématiques dans notre cas car ils conduiraient à construire des requêtes résultant en trop peu de pages Web ou en des pages décorrélées du thème réel du segment. Ainsi, nous avons apporté quelques modifications au critère *tf-idf* standard.

5.1.2 Modifications du critère standard

Pour prendre en considération les spécificités des transcriptions, les scores $S(\ell)$ sont modifiés de diverses manières. Tout d'abord, nous pénalisons le score des noms propres, puis nous prenons en compte les mesures de confiance fournies par le système de transcription pour biaiser les scores *tf-idf* et éviter d'inclure des erreurs de transcription parmi les mots-clés.

Le cas des noms propres

Comme nous l'avons souligné, le score de certains noms tend à être élevé, ceci s'expliquant par le fait que le critère *tf-idf* attribue un score élevé aux mots qui sont spécifiques à un texte donné. Or, la construction de requêtes à partir de noms propres pourrait conduire à un nombre insuffisant de pages pour construire un corpus d'adaptation et estimer des probabilités n-grammes fiables. Par ailleurs, certains de ces noms propres ne décrivent pas le thème d'un segment (des noms de journalistes, de marques...) ou ne donnent qu'une vision partielle de celui-ci. Par exemple, bien que les noms propres Federer et Nadal soient caractéristiques du tennis, leur utilisation pour construire des requêtes ne conduirait probablement qu'à des pages, certes, en nombre suffisant mais exclusivement consacrées à des matchs entre ces deux joueurs plutôt qu'au tennis en général. Si une possibilité pourrait alors être d'écarter purement et simplement les noms propres de la liste des mots-clés, il est malgré tout entendu que certains d'entre eux peuvent participer efficacement à la caractérisation d'un thème. Par exemple, dans un segment traitant des attentats du 11 septembre, les noms propres Manhattan ou Al-Qaïda aideraient sûrement à récolter un corpus thématique de qualité. Au lieu de supprimer les noms propres, nous préférons par conséquent utiliser une stratégie plus douce qui consiste à appliquer une pénalité $p \in [0, 1]$ à la fréquence d'apparition des noms propres. Plus formellement, le terme *tf* d'un lemme ℓ est

Transcription manuelle	Transcription automatique
du tchador et de la cornette au nom du principe de la laïcité de l'état voilà quelques jours nicolas sarkozy rappelait fermement aux musulmans qu'il n'était pas question de porter le voile pour une photo d'identité rappel à l'ordre qui visiblement a échappé à bernadette chirac à en croire le journal le canard enchaîné l'épouse du président de la république se démène pour qu'une religieuse puisse elle garder son voile juliette rengeval le voile de soeur adalberta est tout ce qu'il y a de strict serré sous le menton il encadre son visage sans permettre de voir un seul de ses cheveux alors quand les services de la préfecture exigent que la mère supérieure de la congrégation des soeurs carmélites de l'enfant jésus de bormes-les-mimosas pose tête nue pour sa photo d'identité la religieuse s'insurge elle porte le voile depuis vingt cinq ans son salut se trouve du côté de la première dame de france bernadette chirac qui serait intervenue avec succès auprès de la préfecture pour que soeur adalberta pose voilée affaire un poil embarrassante en plein débat sur la laïcité début avril le ministre de l'intérieur avait rappelé devant les musulmans et musulmanes du congrès de l'uoif que le port du voile était proscrit sur les photos d'identité nicolas sarkozy s'était fait siffler de quoi relancer la polémique le chef de l'état est même intervenu dans un message rendu public cette semaine pour affirmer sa détermination à combattre le communautarisme contraire aux valeurs de la république s'agissant du foulard porté par les musulmanes jacques chirac prône une solution de sagesse quant au voile catholique hier l'élysée semblait avoir fait voeu de silence	du tchador et de la **corne est** au nom du principe de la laïcité de l'état **on l' a** quelques jours nicolas sarkozy **rappeler** fermement aux musulmans **qui** n'était pas question de porter le voile **en** photo **des entités** un rappel à l'ordre qui visiblement a échappé à bernadette chirac **encore** le journal le canard enchaîné l'épouse du président de la république se **démènent** pour qu'une religieuse **puisqu'**elle garder son voile **est une étrange mal** le voile de **serge adda alberta et** tout ce qu'il y a de **strictes série** sous le menton il encadre son visage sans permettre de voir un seul de ses cheveux alors quand les services de la préfecture exigent que la mère supérieure de la congrégation des soeurs carmélites de l'enfant jésus de **borner mimosa aux quatre mille** pour sa photo d' identité un policier s'insurge elle porte le voile depuis vingt-cinq ans son salut se trouve du côté de la première dame de france bernadette chirac qui serait intervenue avec succès auprès de la préfecture pour que **serge adda le vert impôts voilées** affaire un poil **embarrassant** en plein débat sur la laïcité début avril le ministre de l'intérieur avait rappelé devant les musulmans et **musulmane** du congrès de **lui aussi elf** que le port du voile **été proscrits** sur les photos d'identité nicolas sarkozy s'était fait siffler de quoi relancer la polémique le chef de l'état **et** même intervenu dans un message rendu public cette semaine pour affirmer sa détermination à combattre le communautarisme contraire aux valeurs de la république s'agissant du foulard porté par les musulmanes jacques chirac prône une solution de sagesse quant **aux voix catholiques** hier l'élysée semblait avoir fait **le** silence

TABLE 5.1 – Exemple de segment thématique. Les zones en gras indiquent, dans la transcription automatique, les portions de texte mal transcrites.

$S(\ell)$	Mot le plus fréquent	Classe d'équivalence ℓ
1,000	voile	{voile}
0,756	**adda**	{adda}
0,521	**bernadette**	{bernadette}
0,501	laïcité	{laïcité}
0,483	musulmans	{musulmans, musulmane}
0,449	photo	{photo, photos}
0,429	**sarkozy**	{sarkozy}
0,387	**chirac**	{chirac}
0,372	préfecture	{préfecture}
0,364	**serge**	{serge}

TABLE 5.2 – Liste des 10 mots-clés présentant le meilleur score $S(\ell)$ pour le segment de la table 5.1. En gras, les mots trop spécifiques pour la description du thème du segment.

recalculé par :

$$tf'(\ell, t) = \frac{\sum_{\substack{w \in t \\ w \to \ell}} p_w}{|\ell|_t} \times tf(\ell, t) \qquad (5.4)$$

$$\text{avec } p_w = \begin{cases} 1 - p & \text{si } w \text{ est un nom propre,} \\ 1 & \text{sinon,} \end{cases}$$

où $w \to \ell$ signifie que l'occurrence du mot w a pour lemme ℓ et $|\ell|_t$ désigne le nombre de mots de la transcription t qui ont ℓ pour lemme. Un nouveau score $S'(\ell)$ est ainsi obtenu en utilisant tf' au lieu de tf dans la formule (5.3). En pratique, comme nos textes transcrits sont en minuscules et que les noms propres qu'ils contiennent ne sont donc pas caractérisables par ce trait graphique, nous répérons les noms propres grâce à un étiqueteur morphosyntaxique[25] développé pour le traitement de transcriptions (Huet, 2007).

La table 5.3 présente l'effet de cette pénalisation sur les mots-clés précédemment exposés à la table 5.2 en utilisant la valeur p empiriquement fixée à $0,25$. Il apparaît clairement que les noms propres (en gras) sont beaucoup moins présents parmi les premiers mots-clés. Cependant, ces résultats ne sont toujours pas parfaits. D'une part, certains noms propres ne sont pas détectés comme tels car, en l'absence de casse, certaines ambiguïtés sémantiques apparaissent. Par exemple, le mot mimosa peut faire référence à un arbre ou à un terme culinaire alors que, dans le contexte du segment, il réfère le nom propre Bormes-les-mimosas. Bien que les étiqueteurs morphosyntaxiques visent à résoudre automatiquement ce genre d'ambiguïté, ils restent malgré tout faillibles, qui plus est quand, comme c'est le cas ici, des erreurs de transcription viennent se glisser dans le texte. D'autre part, certains noms propres ont par défaut un score tellement élevé que la pénalité p n'a pas de réel impact. Par exemple, le mot de la transcription automatique adda, qui lui aussi résulte d'une erreur de transcription, est toujours positionné deuxième dans notre liste. Ceci pourrait tendre à souligner que la solution idéale serait alors de ne conserver dans nos listes que des noms communs, adjectifs et verbes non modaux. Cependant, comme nous l'évoquions plus haut, cette solution est trop restrictive car certains noms propres peuvent être informatifs pour la caractérisation du thème. Pour dépasser ce problème, la solution pourrait être de généraliser le mécanisme de pénalisation à l'ensemble des catégories grammaticales de

[25] C'est-à-dire un programme qui associe à chaque mot sa classe morphosyntaxique.

	$S'(\ell)$	Mot le plus fréquent	Classe d'équivalence ℓ
–	1,000	voile	{voile}
–	0,567	**adda**	{adda}
△	0,501	laïcité	{laïcité}
△	0,483	musulmans	{musulmans, musulmane}
△	0,449	photo	{photo, photos}
▼	0,391	**bernadette**	{bernadette}
△	0,372	préfecture	{préfecture}
△	0,330	**mimosa**	{mimosa}
△	0,329	tchador	{tchador}
△	0,326	carmélites	{carmélites}
▼	0,322	**sarkozy**	{sarkozy}
▼	0,290	**chirac**	{chirac}
▼	0,273	**serge**	{serge}

TABLE 5.3 – Liste des 10 mots-clés pour le segment de la table 5.1 présentant le meilleur score après pénalisation des noms propres.

manière à privilégier certains types de mots-clés par rapport à d'autres. Cette solution n'est toutefois pas très réaliste car elle nécessiterait une optimisation des différentes pénalités alors qu'il n'existe aucun critère explicite pour comparer la qualité de deux mots-clés, mis à part peut-être, dans notre cadre, la mesure de taux d'erreur après adaptation du sytème de reconnaissance.

Prise en compte des mesures de confiance

Des transcriptions automatiques contiennent inévitablement des erreurs dont l'impact potentiel sur notre processus d'adaptation doit être considéré avec attention. Ces erreurs de transcription peuvent en effet se traduire par des scores biaisés pour certains lemmes et par des listes de mots-clés décorrélées du thème réel d'un segment. Pour amoindrir ces effets, nous proposons de nous appuyer sur les mesures de confiance retournées par le système de reconnaissance pour chaque occurrence de mot dans ses transcriptions automatiques. Plus précisément, nous modifions le score $S'(\ell)$ de chaque lemme ℓ d'après les mesures de confiance des occurrences des mots qui ont pour lemme ℓ, ce qui nous conduit au nouveau score :

$$\sigma(\ell) = \left[\alpha + (1-\alpha)\, c_\ell\right] \times S'(\ell)$$
$$\text{avec} \quad c_\ell = \frac{\sum_{\substack{w \in t \\ w \to \ell}} c_w}{|\ell|_t}, \tag{5.5}$$

où $c_w \in [0,1]$ désigne la mesure de confiance d'un mot w dans la transcription, $w \to \ell$ signifie que w a pour lemme ℓ et $|\ell|_t$ représente le nombre d'occurrences de mots de t qui ont pour lemme ℓ. Le paramètre α, empiriquement fixé à $0,25$, limite l'impact des mesures de confiance, celles-ci n'étant pas parfaitement fiables.

Dans notre exemple, il apparaît que cette prise en compte des mesures de confiance permet de faire reculer les mots adda et mimosa dans la liste des mots-clés alors que les autres mots sont moins affectés par ces pénalités (table 5.4). Néanmoins, nous pouvons remarquer que la liste des premiers mots-clés contient encore le mot adda. Par égard à la complexité de l'ensemble de notre chaîne d'adaptation thématique et à son nombre de paramètres

	$\sigma(\ell)$	Mot le plus fréquent	Classe d'équivalence ℓ
–	0,992	voile	{voile}
△	0,500	laïcité	{laïcité}
△	0,458	musulmans	{musulmans, musulmane}
▼	0,454	**adda**	{adda}
–	0,428	photo	{photo, photos}
–	0,390	**bernadette**	{bernadette}
–	0,371	préfecture	{préfecture}
△	0,328	tchador	{tchador}
△	0,325	carmélites	{carmélites}
△	0,321	**sarkozy**	{sarkozy}
▼	0,294	**mimosa**	{mimosa}

TABLE 5.4 – Listes des 10 mots-clés présentant le meilleur score pour le segment de la table 5.1 après prise en compte des mesures de confiance.

intrinsèques, nous n'avons pas cherché à optimiser l'utilisation des mesures de confiance. Il serait toutefois judicieux d'étudier s'il existe un lien entre le coefficient limitatif α et les mesures de fiabilité des mesures de confiance de notre système (*cf.* section 1.4).

Comme le montre d'autres exemples en annexe B.2, le critère *tf-idf* modifié que nous proposons conduit en moyenne à des listes de mots dont les premiers éléments caractérisent bien la plupart des aspects du thème de leur segment, même si cet avis relève d'un jugement subjectif étant donné qu'aucun critère d'appréciation objectif ne peut être défini. Ces quelques premiers mots de chaque liste peuvent ainsi être considérés comme des mots-clés valables pour formuler des requêtes.

5.2 Formulation de requêtes

Considérant une liste triée de mots-clés, le problème est maintenant de construire une ou plusieurs requêtes qui permettront de collecter des pages Web en rapport avec le thème du segment considéré. Pour cela, deux considérations s'opposent : d'une part, les requêtes doivent être suffisamment précises pour retourner des pages pertinentes pour le thème et, d'autre part, elles ne doivent pas être trop spécifiques, auquel cas trop peu de pages Web seraient renvoyées comme résultat par le moteur de recherche. La recherche d'un compromis passe par un choix judicieux des mots-clés dans les requêtes et par un mécanisme de combinaison de ces mots-clés robuste en terme de nombre et de qualité des pages retournées. En pratique, comme le montre un travail préliminaire rapporté dans (Lecorvé, 2007), il est préférable de ne pas s'appuyer sur des mots-clés classés trop loin dans les listes de chaque segment car la qualité des mots-clés tend à se dégrader avec leur rang, notamment pour les segments les plus courts. Il est souhaitable de ne pas formuler des requêtes contenant trop de mots-clés car le nombre de pages retournées en moyenne pour une requête décroît rapidement lorsque le nombre de ses mots-clés s'accroît.

Pour chaque segment, nous choisissons donc arbitrairement de nous appuyer seulement sur les 5 premiers éléments de la liste de mots-clés et décidons de construire plusieurs requêtes constituées de seulement quelques-uns de ces mots – ce mécanisme d'échantillonage d'un corpus par de multiples requêtes se rapprochant de certaines stratégies du domaine des bases de données, connues sous le nom de *query-based sampling* (Callan et Connell, 2001).

voile	laïcité			
voile		musulmans		
voile			**adda**	
voile				photo
	laïcité	musulmans		
	laïcité		**adda**	
	laïcité			photos
		musulmans	**adda**	
		musulmans		photo
			adda	**photo**

voile	laïcité	musulmans		
voile	**laïcité**		**adda**	
voile	laïcité			photo
voile		**musulmans**	**adda**	
voile		musulmans		photo
voile			**adda**	**photo**
	laïcité	**musulmans**	**adda**	
	laïcité	musulmans		photo
	laïcité		**adda**	**photo**
		musulmans	**adda**	**photo**

TABLE 5.5 – Exemples de 20 requêtes formées à partir de sous-ensembles des 5 mots-clés ayant les meilleurs scores pour le segment de la table 5.1. Les requêtes en gras contiennent des mots mal transcrits.

Par exemple, une première requête est construite à partir des deux meilleurs mots-clés alors qu'une deuxième est constituée du premier et du troisième mots-clés... Après différents essais, il s'avère important de ne pas construire de requêtes à partir d'un seul mot-clé car celui-ci peut être ambigu et biaiser la construction du corpus thématique. Ainsi, dans la version actuelle de notre procédure de création de corpora thématiques, nous construisons pour chaque segment 20 requêtes de 2 ou 3 mots-clés selon un mode de combinaison constant. Comme le présente l'exemple de la table 5.5, cette stratégie originale présente l'avantage d'être assez robuste à l'inclusion de quelques mots erronés parmi les mots-clés. Dans l'exemple, le mot adda n'est en effet présent que dans la moitié des requêtes, il reste donc encore 10 requêtes pertinentes pour construire le corpus thématique.

Cette méthode pourrait probablement être améliorée. Un problème intéressant à résoudre serait de savoir déterminer à l'avance quels couples ou triplets de mots-clés sont pertinents pour former une requête. Il s'agirait notamment de ne plus mélanger entre eux que les seuls 5 meilleurs mots-clés, mais de permettre la combinaison de l'un de ces mots avec un mot-clé au score plus faible à condition que ceux-ci partagent un lien sémantique. Par exemple, dans la table 5.4, les mots-clés voile et tchador partagent une relation d'hyperonymie puisque le tchador est un type de voile. La prise en compte simultanée de ces deux mots-clés au sein d'une requête pourrait peut-être améliorer la pertinence des requêtes en désambiguïsant le mot voile. Cette problématique s'apparente en un sens à la tâche d'extension de requête déjà bien connue en recherche d'information (Claveau *et al.*, 2004 ; Moreau et Claveau, 2006).

5.3 Filtrage thématique de pages Web

Sur la base de la stratégie de formulation de requêtes que nous venons d'exposer, le nombre de pages retournées par les moteurs de recherche atteint fréquemment plusieurs millions. De manière évidente, il n'est pas possible de consulter et de rappatrier l'ensemble de ces pages. Par ailleurs, toutes ces pages ne sont probablement pas pertinentes par rapport au thème du segment, ne serait ce que, comme nous venons de le souligner, parce que certaines requêtes peuvent contenir des mots-clés erronés. Nous avons donc développé une stratégie de filtrage thématique des pages Web. Cette stratégie est explicitée par l'algorithme 5.1. Après avoir lancé chaque requête construite à partir des mots-clés et recensé les

adresses des pages retournées comme résultat, l'agorithme cherche à construire un corpus d'une taille fixée à N documents. Pour cela, il rapatrie à tour de rôle les différentes pages correspondant à chaque requête et mesure la similarité thématique de chacune d'entre elles avec le segment considéré. Selon ce principe, seules les pages dépassant un certain seuil de similarité Δ sont incorporées dans le corpus thématique. Cet algorithme met en avant deux facteurs essentiels : la taille N des corpora et la mise en place d'un critère de similarité thématique. Nous avons étudié chacun de ces aspects.

5.3.1 Taille des corpora

Nous avons mené différentes expériences préliminaires pour déterminer le nombre N de pages à inclure dans les corpora thématiques. Pour cela, nous avons construit un modèle de langue adapté pour chaque segment thématique de notre ensemble de développement en interpolant linéairement notre modèle de langue généraliste avec un modèle de langue d'adaptation appris à partir de chaque corpus thématique obtenu sans filtrage thématique. Nous mesurons alors la variation de la perplexité de ces modèles par rapport à celle du modèle généraliste pour différentes tailles de corpus thématique. Ces résultats sont présentés à la figure 5.2. Il apparaît que l'augmentation du nombre de pages par corpus thématique produit des diminutions de la perplexité des modèles adaptés par rapport à celle du modèle de langue généraliste. Ceci est toutefois à pondérer par le fait que, à partir de 100 pages, l'intérêt d'ajouter de nouvelles pages tend à s'amoindrir petit à petit. Ce constat peut vraisemblablement s'expliquer par le fait que, au fur et à mesure que l'on avance dans les listes de résultats fournis pour chaque requête, la pertinence des pages par rapport à une requête se dégrade et leur intérêt devient donc moindre pour notre tâche d'adaptation thématique. Par ailleurs, comme le montre la figure 5.3, il est difficile de faire un rapprochement direct entre la taille en mots des corpora thématiques et les gains relatifs en terme de perplexité. Ainsi, dans la suite des expériences, nous avons estimé que des corpora de 200 pages étaient suffisants et constituaient un bon compromis entre qualité et vitesse de construction des corpora thématiques. Ce nombre de pages aboutit à considérer des corpora thématiques d'environ 800 000 mots.

5.3.2 Filtrage thématique

Parmi l'ensemble des pages renvoyées par les requêtes liées à un segment, il n'est pas rare de rencontrer des pages linguistiquement trop pauvres ou thématiquement éloignées du segment. Ainsi, nous avons proposé une mesure de similarité thématique pour les filtrer.

Avant toute chose, cette étape de filtrage nécessite d'accéder au contenu principal d'une page Web. Or, celui-ci est généralement noyé dans une quantité d'informations inintéressantes comme des menus, des mentions légales, des publicités... Nous avons donc développé une méthode de nettoyage d'une page HTML qui supprime les balises de mise en forme mais aussi écarte les zones de texte ne correspondant pas au contenu central de la page. Comme le détaille l'annexe C, ce nettoyage s'appuie principalement sur un élagage de l'arbre DOM d'un document HTML à partir de critères statistiques sur des indices tels que la fréquences des signes de ponctuation, la taille des phrases, le nombre de caractères non alpha-numériques... Alors que peu de méthodes existaient pour le nettoyage de pages Web au moment de nos travaux, il est bon de noter que cette tâche a depuis fait l'objet d'un gain d'intérêt et que des techniques plus sophistiquées existent donc aujourd'hui (Baroni *et al.*, 2008).

```
Fonction construire_corpus( R : requêtes,  t : transcription) : corpus thématique
    url : tableau de listes d'URL ;
    u : URL ;
    page : texte d'une page ;
    corpus : ensemble de textes ;
    r : requête courante ;

    // Lancement de toutes les requêtes et stockage des URL retournées
    Pour chaque requête r de R faire
        url[rq] ← lancer_requête(r) ;
    Fait

    // Sélection de N pages pertinentes
    corpus ← ∅ ;
    r ← première_requête(R) ;
    Tant que |corpus| < N faire
        u ← prochaine_url(url[r]) ;
        p ← rapatrier(u) ;
        p ← nettoyer(p) ;
        Si (similarité_thématique(t, p) ⩾ Δ) Alors
            corpus ← corpus ∪ {p} ;
        Fin Si
        r ← requête_suivante(R) ; // Lorsque la dernière requête est atteinte,
                                  // on retourne à la première requête.
    Fait
    Retourner corpus ;
Fin
```

ALGORITHME 5.1 – Pseudo-code de la construction d'un corpus thématique à partir d'un ensemble de requêtes et d'une transcription automatique.

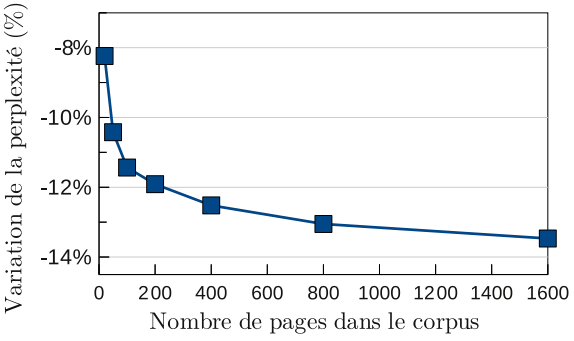


FIGURE 5.2 – Variation relative moyenne de la perplexité en fonction du nombre de pages des corpora thématiques.

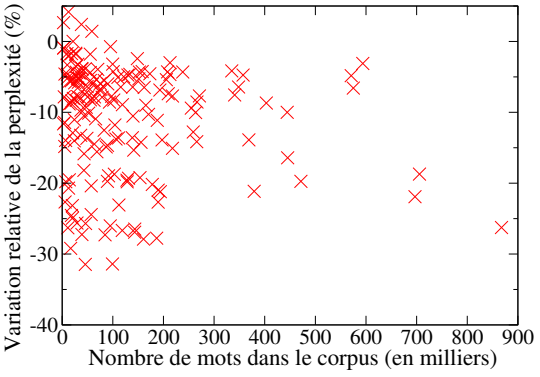


FIGURE 5.3 – Variation relative de la perplexité en fonction du nombre de mots de chaque corpus de 50 ou 100 pages.

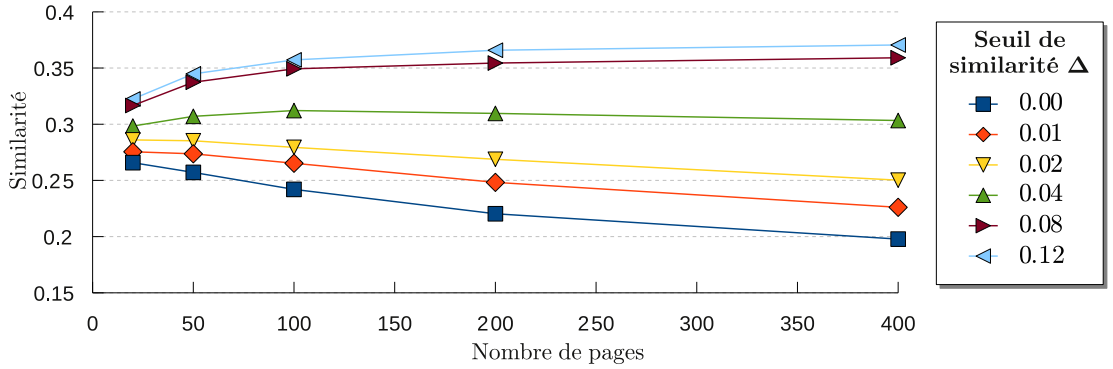


FIGURE 5.4 – Similarité moyenne sur l'ensemble de développement entre un segment et son corpus thématique en fonction du nombre de pages et du seuil de similarité.

Une fois nettoyée, nous mesurons la similarité thématique d'une page p avec la transcription t du segment considéré en nous appuyant à nouveau sur le critère *tf-idf*. En représentant t comme un vecteur de scores $\sigma_t(\ell)$ et p comme un vecteur de scores $S'_p(\ell)$[26], la similarité entre t et p est calculée comme une mesure cosinus entre leur vecteurs respectifs :

$$\mathrm{sim}(t,p) = \frac{\sum\limits_{\ell \in t \cap p} \sigma_t(\ell) \times S'_p(\ell)}{\sqrt{\sum\limits_{\ell \in t} \sigma_t(\ell)^2 \times \sum\limits_{\ell \in p} S'_p(\ell)^2}}. \tag{5.6}$$

Les pages dont la similarité thématique est inférieure à un seuil Δ sont écartées par l'algorithme. La figure 5.4 atteste de l'utilité de ce filtrage thématique en mesurant la similarité globale d'un corpus thématique avec la transcription de son segment pour différentes tailles en nombre de pages et différents seuils de similarité sur l'ensemble de développement. On constate que, sans seuil de similarité ($\Delta = 0$), l'augmentation de la taille des corpora tend à dégrader leur qualité. À l'inverse, cette qualité augmente puis se stabilise lorsqu'un seuil supérieur à $0,08$ est utilisé.

En pratique, pour notre tâche d'adaptation thématique, ce seuil a été déterminé empiriquement en mesurant les variations de la perplexité des modèles adaptés par une technique d'interpolation linéaire sur l'ensemble de développement. La figure 5.5 reporte ces variations pour différents coefficients d'interpolation. Il apparaît alors que, quel que soit le coefficient d'interpolation utilisé pour l'adaptation du modèle généraliste, l'augmentation du seuil de similarité conduit à des perplexités plus basses. On remarque toutefois que l'utilisation d'un seuil de similarité trop élevé conduit à dégrader cette qualité. Ce constat s'explique par le fait que, passé un certain degré de similarité, l'algorithme ne parvient plus à trouver suffisamment de pages, ce qui résulte en des corpora thématiques plus petits conduisant donc à une moins bonne estimation d'un modèle de langue. Ainsi, dans notre processus d'adaptation, nous fixons le seuil de similarité à $\Delta = 0,1$.

[26] Nous n'utilisons pas le score $\sigma(\ell)$ pour une page Web car aucune mesure de confiance sur ses mots n'existe.

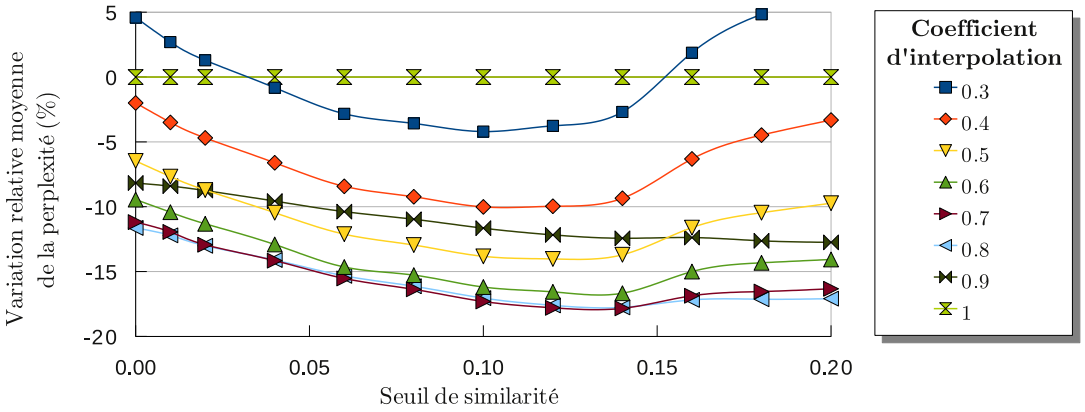


FIGURE 5.5 – Variations relatives moyennes de la perplexité sur l'ensemble de développement pour différents seuils de similarité et différents coefficients d'interpolation pour des corpora de 200 pages. Un coefficient d'interpolation à 1 correspond à l'utilisation du seul modèle de langue généraliste.

5.4 Validation et mise en application

L'étape de création de corpora thématiques s'incrit dans le cadre de notre processus d'adaptation thématique d'un système de reconnaissance automatique de la parole. Cette section vise ainsi à valider la chaîne de traitements que nous venons de décrire avant de poursuivre la présentation des autres étapes dans les chapitres suivants. Pour cela, nous évaluons la qualité des corpora thématiques créés pour la seule adaptation du modèle de langue généraliste de notre système. Plus précisément, nous nous appuyons sur une classique technique d'interpolation linéaire et mesurons le WER obtenus sur nos segments par le système avant et après adaptation. Parallèlement, nous montrons également que le processus que nous avons développé peut être appliqué dans un autre contexte, à savoir celui de la structuration de flux multimédias.

5.4.1 Validation *via* l'adaptation d'un modèle de langue

Pour juger de la qualité des corpora thématiques créés après optimisation des différents paramètres intrinsèques à notre méthode, nous utilisons les modèles adaptés à partir de nos corpora pour chercher à améliorer la transcription automatique de nos segments. Ces modèles sont calculés par l'interpolation linéaire de notre modèle de langue généraliste avec un modèle de langue d'adaptation appris sur le corpus thématique de chaque segment. Cette dernière technique passe par l'optimisation d'un coefficient d'interpolation λ. Pour cela, nous avons mesuré la variation de la perplexité des modèles adaptés par rapport à celle du modèle généraliste pour différentes valeurs du coefficient d'interpolation. Ces résultats sont présentés par la figure 5.6 pour deux jeux de paramètres intéressants de notre processus de création de corpora thématiques. Il apparaît qu'un coefficient situé entre $0,7$ et $0,8$ semble conduire aux meilleures perplexités. On remarque toutefois que l'utilisation d'une valeur λ identique pour tous les segments n'est pas optimale. Les droites en pointillés de cette figure montrent que de meilleures variations pourraient être atteintes si l'on savait ajuster au mieux le coefficient d'interpolation de manière indépendante pour chaque segment. Ceci nécessiterait de savoir diagnostiquer à quel point un corpus thématique peut apporter à la

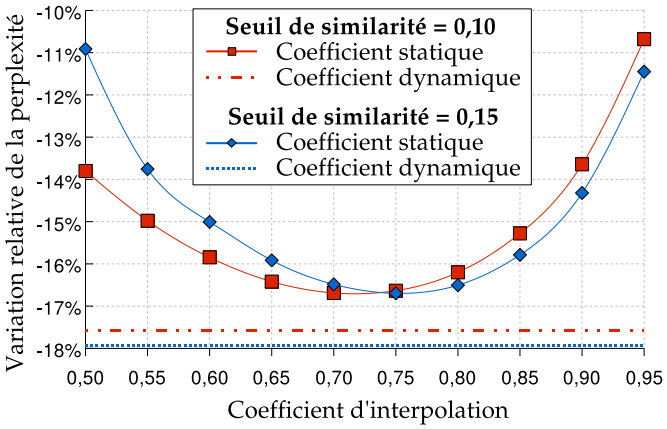


FIGURE 5.6 – Impact du coefficient d'interpolation sur la perplexité selon qu'il soit constant pour tout segment (lignes continues) ou fixé de manière optimale pour chaque segment.

modélisation linguistique du système pour un segment donné. Il pourrait par exemple être intéressant d'étudier s'il existe une corrélation entre la similarité thématique globale d'un corpus avec son segment et la valeur optimale du coefficient d'interpolation pour chaque segment. Il s'agit toutefois d'une question que nous n'avons pas traitée. Ainsi, nous utilisons un coefficient d'interpolation constant $\lambda = 0,8$.

Ce coefficient fixé, nous générons alors de nouvelles transcriptions à partir des modèles interpolés en ré-évaluant les graphes de mots obtenus après la première passe de décodage du système IRENE grâce au modèle de langue généraliste. La table 5.6 compare les différents WER obtenus sur les ensembles de développement et de test. Ces résultats mettent clairement en évidence le fait que, quelle que soit la radio d'où proviennent les segments, les modèles adaptés conduisent à des transcriptions automatiques de meilleure qualité que celles du modèle généraliste. Un gain absolu de $0,2$ est ainsi obtenu sur l'ensemble de test, ce gain étant statistiquement significatif d'après les tests de Student ($p = 7 \times 10^{-5}$) et de Wilcoxon ($p = 1 \times 10^{-5}$). Le détail de ces résultats montrent toutefois que les gains obtenus sur l'ensemble de test sont plus faibles que ceux obtenus pour l'ensemble de développement. Ceci n'est toutefois pas surprenant étant donné que les divers réglages de notre technique ont été effectués sur l'ensemble de développement. Par ailleurs, on remarque que les gains les plus importants sont obtenus pour les segments de la radio RFI. Ceci s'explique sans doute par le fait que les sujets abordés sur cette radio, fréquemment liés à l'Afrique, sont peu présents dans le corpus d'apprentissage du système – pour rappel, principalement des articles issus du journal *Le Monde* – et sont donc mal modélisés par le modèle de langue généraliste de notre système. À l'inverse, ceux de France Inter et France Info traitent, par exemple, plus volontiers de politique nationale et internationale, sujets largement traités dans ce quotidien français. Ainsi, l'effet de l'adaptation thématique se fait en moyenne plus ressentir pour les thèmes abordés sur RFI que pour ceux des deux autres radios. Toujours est-il que les résultats globaux attestent de la qualité et du potentiel des corpora thématiques construits d'après notre procédure pour la tâche d'adaptation thématique.

Par ailleurs, nous avons cherché à savoir si les gains observés venaient bien de l'utilisation d'un corpus thématiquement cohérent et non pas simplement de l'apport de nouvelles données d'apprentissage. Pour cela, nous avons regroupé l'ensemble des corpora thématiques obtenus pour chaque segment de notre ensemble de développement et utilisé ces données pour l'interpolation du modèle de langue généraliste. Après avoir retranscrit chaque

	Modèle de langue	France Inter	France Info	RFI	Global
Ensemble de développement	Généraliste	20,5	21,0	25,6	22,4
	Adapté	20,1 (−0,4)	20,6 (−0,4)	25,1 (−0,5)	21,9 (−0,5)
Ensemble de test	Généraliste	19,9	21,7	23,5	21,7
	Adapté	19,7 (−0,2)	21,6 (−0,1)	23,2 (−0,3)	21,5 (−0,2)

TABLE 5.6 – WER observés sur les ensembles de développement et de test, pour chaque radio et globalement, en utilisant soit le modèle de langue généraliste soit les modèles adaptés. Les variations absolues sont présentées entre parenthèses.

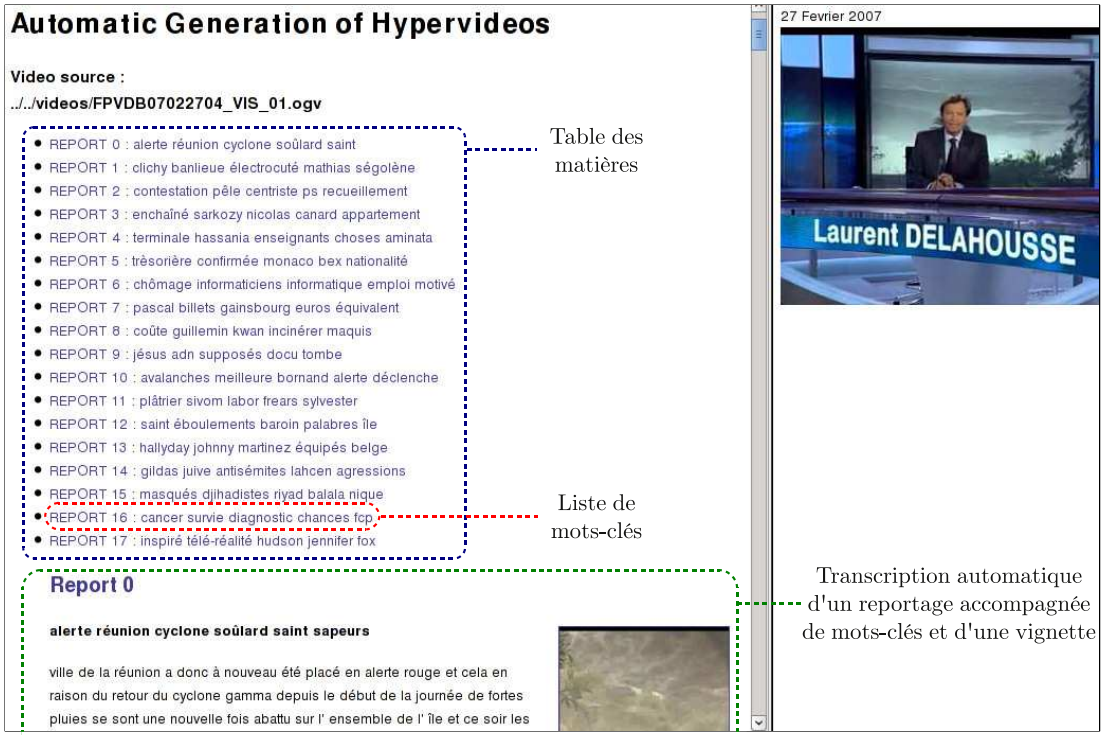
segment avec ce nouveau modèle et remesuré le WER, il n'est alors ressorti qu'un gain absolu moyen de $0,1$, ce qui est nettement inférieur au gain de $0,5$ reporté précédemment lorsque l'interpolation s'effectue bien sur la base d'un corpus spécifiquement dédié à chaque segment.

5.4.2 Application à l'enrichissement de documents multimédias

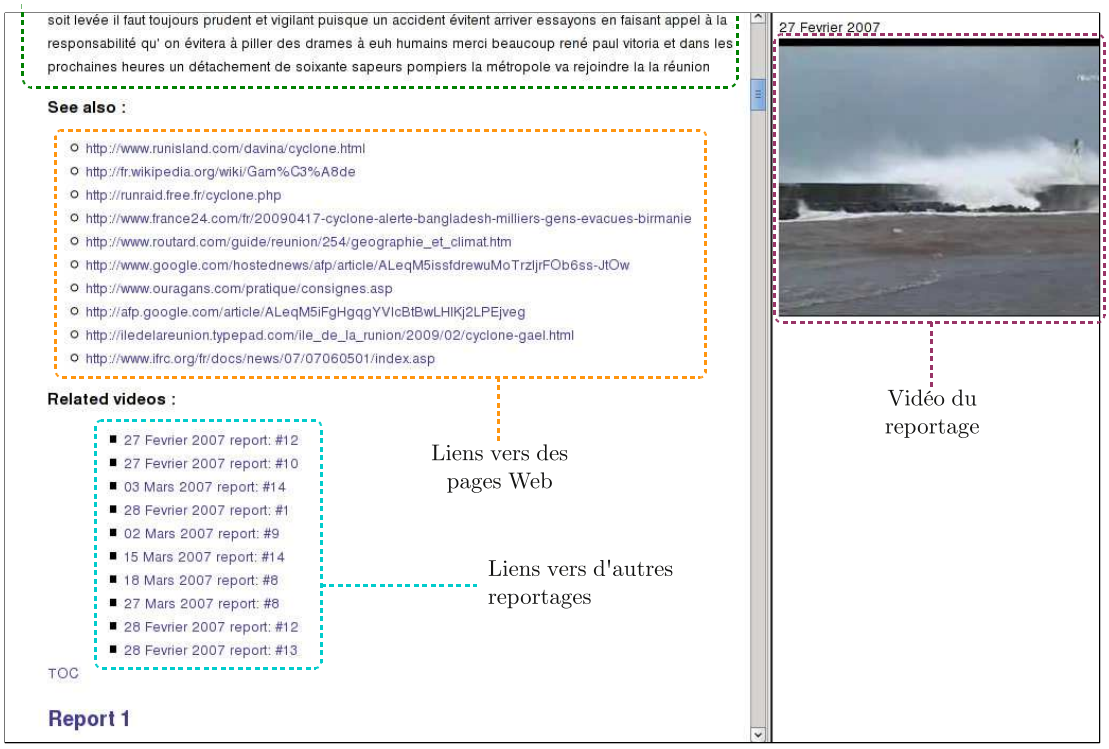
Lors d'un travail plus récent, nous avons montré que notre processus de création de corpora peut aussi servir à enrichir la description de documents multimédias dans une application de délinéarisation de flux télévisuels. Globalement, cette application vise à segmenter automatiquement des flux de télévision en documents indépendants (programmes, reportages...) et à générer automatiquement une description de ces segments pour faciliter leur accès par des utilisateurs. Dans ce cadre, nous avons montré que l'étape d'extraction de mots-clés sur la transcription automatique pouvait permettre de caractériser rapidement le contenu d'un document. De plus, nos mécanismes de recherche sur Internet et de filtrage thématique sont exploitables pour mettre en lien le document avec d'autres, extraits du même flux multimédia, ainsi qu'avec des pages Web en rapport avec le contenu du document. De cette manière, les possibilités de navigation au sein d'une base de documents se trouvent accrues et facilitées.

Cette étude présente quelques particularités par rapport à notre travail précédent. D'une part, comme cette application de délinéarisation est actuellement appliquée à des journaux télévisés, la création de liens entre divers documents consiste à rassembler des documents traitant d'un même fait d'actualité précis plutôt que d'un même thème. Ainsi, des aménagements sont à apporter à la technique présentée dans ce chapitre. Notamment, notre traitement spécifique des noms propres ne se justifie plus autant étant donné que ceux-ci sont des éléments centraux pour décrire un évènement précis. D'autre part, cette tâche représente un défi intéressant puisqu'il s'agit de travailler sur des documents issus d'une segmentation automatique alors que les documents que nous utilisons pour le reste de notre travail sont issus d'une segmentation thématique manuelle. Malgré ces particularités, nous avons pu montrer le bon comportement de notre méthode à travers le développement d'une démonstration présentée lors du *NEM Summit 2009*[27] dont deux captures d'écran illustrent le contenu en figure 5.7.

[27]Nous tenons au passage à remercier Mathieu Ben, Sébastien Campion et Camille Guinaudeau pour leurs contributions respectives au sein ce travail collaboratif.



(a) – Table des matières des segments extraits d'un journal télévisé. Chaque document est décrit par quelques mots-clés.



(b) – Présentation d'un reportage à travers sa vidéo et sa transcription automatique accompagnées de liens vers des pages Web et d'autres reportages liés au même sujet d'actualité.

FIGURE 5.7 – Capture d'écran de notre démonstration de délinéarisation de flux multimédia pour la création de vidéos enrichies.

5.5 Améliorations possibles

Différentes pistes d'amélioration mériteraient d'être étudiées, notamment en ce qui concerne l'étape d'extraction de mots-clés. Celle-ci devrait être rendue plus robuste car elle est cruciale pour le reste du processus.

Tout d'abord, comme nous l'avons vu dans l'exemple déroulé tout au long de ce chapitre, certains mots mal transcrits peuvent rester présents parmi les mots-clés d'un segment, ce en dépit de l'utilisation des mesures de confiance. Pour pallier ce problème, il serait bon d'intégrer des relations sémantiques entre mots pour accorder plus d'importance aux mots-clés qui donnent au segment une cohérence sémantique plus forte. L'utilisation de ce type de relations est d'autant plus envisageable qu'il est possible de les apprendre automatiquement à partir d'un corpus par des techniques qui mesurent la similarité des contextes d'apparition de mots distincts (Grefenstette, 1994). Dans notre exemple, ces relations pourraient ainsi mettre en lumière que le mot adda n'est pas corrélé avec l'ensemble des autres mots-clés et qu'il devrait donc être écarté. Ces relations pourraient également profiter au filtrage thématique des pages Web en enrichissant les vecteurs de scores *tf-idf* utilisés comme représentation des documents. Par exemple, la description d'une page comportant le mot vélo pourrait être complétée par l'attribution d'un score au mot bicyclette même si ce mot n'apparaît pas dans la page. Ceci permettrait probablement de rendre l'étape de filtrage des pages Web, d'une part, plus souple pour des pages thématiquement pertinentes mais relevant d'un registre de langue différent et, d'autre part, plus stricte pour des pages ne partageant fortuitement que certains mots avec la transcription automatique d'un segment.

Ensuite, il pourrait s'avérer judicieux de chercher à dissocier différents aspects du thème d'un document, notamment lorsque l'algorithme ne parvient pas à récolter suffisamment de pages Web pertinentes. Par exemple, si un segment traite du procès de *Saddam Hussein*, il pourrait être bon de dissocier automatiquement les mots-clés liés à l'Irak de ceux rattachés au domaine de la justice. Pour réaliser cette opération sans s'appuyer sur une classification *a priori* des thèmes, il serait par exemple envisageable de récolter un premier corpus thématique de manière classique et d'effectuer un regroupement automatique de ces pages d'après notre mesure de similarité thématique avant de lancer la récupération de corpora thématiques liés à chaque groupe de pages. Sans réaliser de *clustering*, ce processus serait d'ailleurs également intéressant pour tenter de caractériser plus finement le thème d'un segment. Toutefois, ce genre de processus itératif serait coûteux en temps, et, ce qui est plus gênant, pourrait petit à petit faire dévier le thème des corpora récupérés pour finalement converger vers un thème tout autre que celui du segment considéré.

Enfin, il serait intéressant de remplacer ou compléter la notion de mots-clés en considérant des séquences de mots plutôt que des mots isolés. En guise d'illustration, remarquons que, dans notre exemple, les mots photo et identité sont beaucoup plus informatifs pour la description du segment lorsqu'il sont considérés ensemble plutôt que séparément. Pour cela, il est envisageable de grouper certains mots-clés via des techniques d'extraction de séquences complexes (Manning et Schütze, 1999 ; Daille, 2002). Toutefois, ces possibilités risquent d'être limitées par la difficulté à comparer, au sein du critère *tf-idf*, les scores d'unités de granularités différentes, comme des mots et des séquences de mots. Typiquement, cette difficulté tient dans le fait que la fréquence moyenne d'apparition d'une séquence de mots est plus faible que celle des mots qui la composent et que ces structures complexes sont fortement sujettes à variation.

5.6 Bilan du chapitre

Dans ce chapitre, nous avons exposé notre technique originale de création d'un corpus thématique à partir de la transcription automatique d'un segment sans utiliser de données *a priori* sur les thèmes rencontrés ni recourir à des traitements manuels. Plus précisément, nous avons défini un processus de caractérisation du thème d'un segment en aménageant le critère *tf-idf*. Cette caractérisation nous permet d'extraire des mots-clés qui, une fois combinés au sein de multiples petites requêtes, aboutissent à la récupération de pages Web. Nous avons également proposé un mécanisme original de filtrage thématique de ces pages qui conduit à améliorer la qualité des corpora thématiques pour modéliser la langue au sein du thème d'un segment. Ces corpora nous permettent déjà d'obtenir des gains en terme de taux de reconnaissance par l'utilisation d'une technique d'interpolation linéaire. À part l'optimisation du coefficient d'interpolation de cette technique, nous n'avons pour l'instant pas cherché à utiliser une méthode optimale de réestimation du modèle de langue. Plus particulièrement, aucun n'effort n'a été porté quant à la prise en compte des spécificités linguistiques de chaque thème rencontré. Le chapitre suivant s'intéresse ainsi à la mise en place d'une telle adaptation thématique en s'appuyant sur des corpora thématiques créés *via* notre technique.

Chapitre 6

Adaptation thématique d'un modèle de langue

Ce chapitre poursuit l'étude des différentes étapes de notre processus d'adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole en abordant la tâche d'adaptation d'un modèle de langue généraliste. Sur un plan général, cette tâche a pour objectif de rendre plus probables les séquences de mots spécifiques à un thème considéré. Dans notre contexte, il s'agit d'utiliser un corpus thématique construit automatiquement à partir d'Internet pour réestimer les probabilités n-grammes de notre modèle initial de manière à mieux prendre en compte ces spécificités. Cette réestimation doit permettre d'obtenir une meilleure transcription automatique d'un segment du même thème, notamment au niveau des mots importants pour le thème considéré. Différentes tâches se dessinent au sein de cette problématique. Tout d'abord, il faut être en mesure de déterminer à partir d'un corpus quelles sont ces probabilités n-grammes à modifier. Ensuite, il convient de savoir comment mettre à jour ces probabilités. Enfin, une fois le modèle de langue adapté, nous devons être en mesure d'en faire le meilleur usage afin d'optimiser le taux de reconnaissance de notre système de transcription.

Nous commençons dans ce chapitre par répondre à cette dernière question en reprenant la méthode d'interpolation linéaire présentée au chapitre précédent et en examinant à quelle étape du processus de transcription les modèles ainsi adaptés permettent d'obtenir des performances optimales en terme de WER. Ceci nous permet de fixer un cadre d'évaluation commun à toutes les expériences de ce chapitre et nous conduit dans un même temps à souligner les défauts de l'interpolation linéaire. Ensuite, nous répondons aux deux autres questions en présentant nos contributions pour l'adaptation thématique d'un modèle de langue. Plus précisément, nous introduisons une méthode d'adaptation par minimum d'information discriminante (MDI) fondée sur l'utilisation de terminologies construites automatiquement à partir des corpora thématiques. En mettant en lumière les mots et séquences de mots spécifiques à un thème, respectivement par l'acquisition de *termes simples* et des *termes complexes*, ces terminologies nous permettent en effet de restreindre la réestimation des probabilités d'un modèle de langue à un sous-ensemble de n-grammes ayant un rôle important dans l'usage de la langue au sein du thème considéré. La présentation de cette stratégie d'adaptation est découpée en trois sections. La section 6.2 donne une description de l'adaptation MDI dans un cadre générique et introduit les notions principales du problème d'acquisition d'une terminologie à partir d'un corpus spécialisé. Les

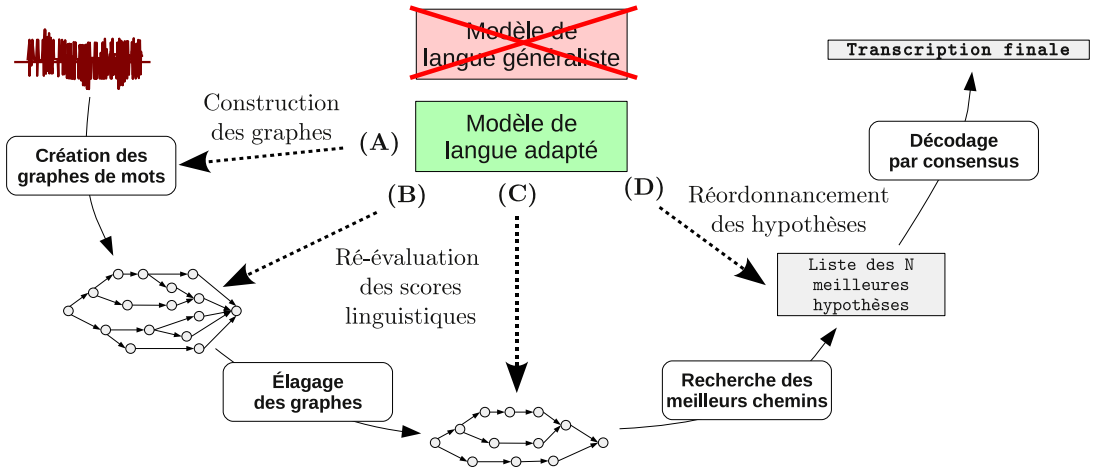


FIGURE 6.1 – Schéma des différentes étapes du processus de transcription à partir desquelles un modèle de langue adapté peut être intégré après un premier décodage complet par un modèle de langue généraliste.

sections 6.3 et 6.4 présentent alors différentes instanciations de l'adaptation MDI fondées respectivement sur l'utilisation de termes simples et sur celle de termes complexes. Enfin, la section 6.5 conclue ce chapitre en diagnostiquant les limites de notre méthode et, plus généralement, celles de l'adaptation d'un modèle de langue n-gramme.

Deux précisions importantes sont à donner quant aux expériences menées pour ces différentes sections. D'une part, celles-ci s'appuient toutes sur les mêmes corpora thématiques récupérés et stockés une fois pour toutes tel que nous l'avons décrit au chapitre précédent. D'autre part, pour une meilleure intelligibilité des résultats concernant la réestimation de probabilités existantes dans un modèle de langue, aucune modification du vocabulaire n'est faite et toutes les expériences se fondent donc sur le vocabulaire initial du système.

6.1 Retour sur l'interpolation linéaire

La technique la plus immédiate pour adapter notre modèle de langue généraliste consiste à interpoler ce dernier avec un modèle de langue d'adaptation appris sur un corpus thématique. Nous avons montré au chapitre précédent que, après optimisation du coefficient d'interpolation, cette méthode aboutit à des gains sur le WER lorsque les modèles ainsi adaptés sont utilisés pour ré-évaluer les graphes de mots issus de la première passe du processus de transcription ayant permis d'obtenir une transcription initiale du segment considéré (*cf.* le processus de transcription du système IRENE à la section 1.5).

Dans cette section, nous étudions la pertinence de ce mode d'utilisation des modèles adaptés et analysons l'impact de ces modèles sur les transcriptions qu'ils permettent de générer.

6.1.1 Intégration du modèle adapté dans le processus de transcription

Comme le montre la figure 6.1, l'intégration d'un modèle de langue adapté peut se faire en reprenant à différents niveaux le processus de transcription préalablement effectué avec le modèle généraliste. Tout d'abord, un nouveau modèle peut être directement utilisé pour

	Sans adaptation	Intégration à partir de A	B	C
Ensemble de développement	22,4	21,7 $(-0,7)$	21,9 $(-0,5)$	22,0 $(-0,4)$
Ensemble de test	21,7	21,2 $(-0,5)$	21,5 $(-0,2)$	21,7 $(0,0)$

TABLE 6.1 – WER mesurés sans adaptation thématique et après adaptation et intégration du modèle de langue à partir de différentes étapes du processus de transcription.

relancer un processus complet de transcription (légende A). Ensuite, comme nous le faisons jusqu'à présent, ce modèle peut n'être intégré qu'après la création des graphes de mots par le modèle de langue généraliste et servir alors à ré-évaluer les scores linguistiques de ces graphes. Il peut s'agir des graphes de mots relativement denses résultant de l'algorithme de décodage en faisceau (légende B) ou de ceux plus pauvres en hypothèses issus d'un élagage ultérieur (légende C). Enfin, l'utilisation d'un modèle de langue adapté peut n'être envisagée que pour réordonner les N meilleures hypothèses générées précédemment par le modèle généraliste (légende D).

La table 6.1 présente les taux d'erreur obtenus sur nos ensembles de développement et de test en intégrant les modèles de langue adaptés à partir de différents étapes à l'exception de celle de réordonnancement. En effet, cette étape ne nous intéresse pas car notre système possède déjà une passe de post-traitement des listes des $\mathcal{N}$ meilleures hypothèses (*cf.* section 1.5) et nous ne voulons pas interférer avec ce traitement. Comme le montre les différents WER obtenus, l'apport des modèles adaptés est d'autant plus marqué qu'ils sont intégrés tôt dans le processus de transcription. Ainsi, il apparaît qu'il est préférable de relancer un décodage complet avec les nouveaux modèles si l'on cherche à obtenir les meilleures améliorations. Ce constat n'a rien d'étonnant étant donné que cette solution permet aux nouveaux modèles de se confronter à l'espace de recherche entier alors que les autres configurations imposent à ces modèles de se restreindre aux seules hypothèses que le modèle généraliste avait précédemment jugées pertinentes. Ainsi, sur l'ensemble de test, on note que, en intégrant les modèles adaptés à l'étape C, ces derniers ne parviennent pas à produire de meilleures hypothèses que celles données par le modèle généraliste alors qu'un gain absolu de $0,5$ est obtenu lorsque les modèles sont intégrés au plus tôt.

Ces résultats nous conduisent, pour le reste de nos travaux, à intégrer systématiquement les modèles de langue adaptés dès l'étape de création des graphes de mots. Par ailleurs, ces expériences nous donnent l'occasion d'analyser l'impact, sur les sorties du système, de modèles de langue thématiquement adaptés par une technique d'interpolation linéaire.

6.1.2 Analyse des transcriptions après interpolation linéaire

L'interpolation linéaire étant une méthode classique qui peut s'appliquer à n'importe quelle tâche d'adaptation, nous avons souhaité savoir en quoi cette technique permet d'améliorer les transcriptions de nos segments. Pour illustrer cette analyse, nous présentons dans la table 6.2 trois exemples symptomatiques de ce que nous pouvons observer en comparant les transcriptions obtenues grâce au modèle de langue généraliste et grâce à un modèle interpolé à partir d'un corpus thématique.

Thème : laïcité, port du voile	
Référence	tête nue
Modèle généraliste	SEPT MILLE
Modèle adapté	tête nue
Thème : grève des transports en commun	
Référence	le service des trams est affecté
Modèle généraliste	le service des TRAMES est affectée
Modèle adapté	le service des TRAM est affectée
Thème : Irak, États-Unis, soldats	
Référence	cent vingt trois dont une grande partie
Modèle généraliste	cent vingt trois dont une grande partie
Modèle adapté	cent vingt trois DANS une grande partie

TABLE 6.2 – Comparaison de groupes de souffle transcrits avec le modèle généraliste ou avec un modèle adapté par rapport à la transcription de référence pour des segments traitant de différents thèmes. Les mots en majuscules sont les mots mal transcrits.

	Sans adaptation	Intégration à partir de A	B	C
Ensemble de développement	20,1	19,2 (−0,9)	19,3 (−0,8)	19,2 (−0,9)
Ensemble de test	19,6	18,7 (−0,9)	19,1 (−0,5)	20,0 (+0,4)

TABLE 6.3 – LER mesurés sans adaptation thématique et après adaptation et intégration du modèle de langue à partir de différentes étapes du processus de transcription.

Globalement, la confrontation des sorties fournies par ces deux configurations montre que les modèles interpolés permettent de corriger certaines erreurs sur les mots importants pour le thème considéré. Toutefois, il est possible de constater que, dans certains autres cas, cette correction n'est que partiellement opérée puisque les modèles adaptés corrigent le lemme de certains mots thématiquement importants précédemment mal transcrits mais ne parviennent pas à produire en sortie la bonne flexion de ces mots. Ce phénomène s'explique par les disparités qu'il existe en terme de fréquence d'apparition entre flexions d'un même lemme dans les corpora thématiques. Il en découle que les probabilités n-grammes liées à certaines flexions augmentent, alors que d'autres restent inchangées, voire diminuent sous l'effet de l'interpolation linéaire des probabilités estimées à partir des corpora d'adaptation avec celles préalables du modèle généraliste. Par ailleurs, ce problème d'éparsité s'applique également à la réestimation des probabilités de n-grammes correspondant à des séquences de mots du langage général. Il en découle des dégradations dans les transcriptions sur des mots ordinaires (verbes auxiliaires ou de modalité, prépositions, déterminants...).

De manière générale, ces insuffisances sont confirmées par le calcul d'un taux d'erreur sur les lemmes des mots lexicaux (LER), ce taux permettant de faire abstraction des erreurs de flexion et des erreurs sur les mots vides. Les résultats de la table 6.3 nous montrent en effet que les gains reportés pour le LER sont supérieurs à ceux présentés précédemment sur le WER. Ceci s'explique par le fait que l'interpolation linéaire se limite aux séquences de mots observées dans le corpus thématique, et n'adapte donc que partiellement le modèle

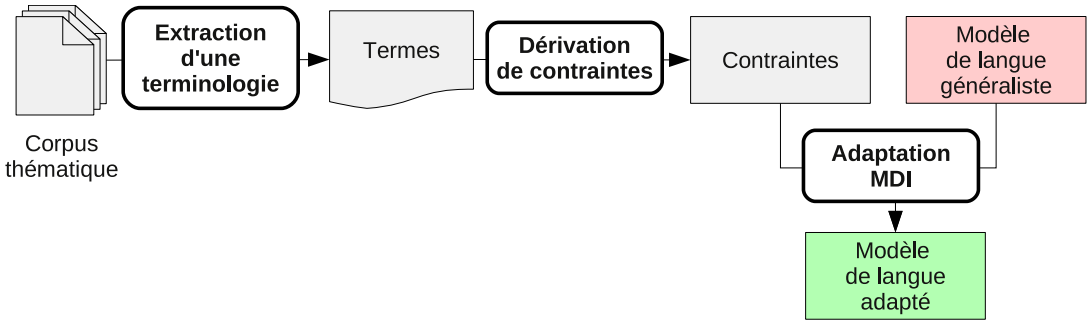


FIGURE 6.2 – Schéma de notre processus d'adaptation d'un modèle de langue fondé sur l'extraction d'une terminologie et sur l'utilisation de MDI.

de langue généraliste en ne s'intéressant qu'à certaines flexions et en ne modifiant que les probabilités de certains n-grammes contenant ces flexions. Notre but est donc de développer une adaptation plus cohérente qui exploite plus intelligemment le contenu des corpora thématiques, notamment par des techniques de traitement automatique des langues.

Après une analyse des expériences préliminaires sur l'adaptation d'un modèle de langue, deux constats émergent donc. D'une part, l'intégration des modèles adaptés doit être faite au plus tôt. Ainsi, dans le reste de nos expériences, nous intégrons les modèles adaptés dès la phase de création des graphes. D'autre part, il apparait que les corpora thématiques possèdent un potentiel que n'exploite qu'en partie la technique d'interpolation linéaire car cette dernière adapte de la même manière tous les n-grammes et n'intègre aucune connaissance linguistique. Nous proposons alors une méthode plus élaborée où les objectifs de l'adaptation thématique sont explicités en déterminant automatiquement les n-grammes dont il faut augmenter les probabilités à partir de techniques de traitement automatique des langues.

6.2 Adaptation MDI à partir d'une terminologie

L'interpolation linéaire a l'inconvénient d'adapter les probabilités n-grammes d'un modèle de langue indépendamment de leur importance au sein du thème considéré et de n'intégrer aucune connaissance linguistique, ce qui aboutit à une mauvaise réestimation de certaines probabilités. Ce problème vient du fait que nos corpora thématiques sont, certes, représentatifs du langage employé au sein d'un thème mais que, d'une part, ceux-ci ne nous permettent en rien de diagnostiquer quels autres n-grammes sont, au contraire, peu enclins à apparaître au sein de ce même thème et que, d'autre part, ils souffrent d'un problème d'éparsité des données. Pour pallier ce problème, nous proposons une méthode d'adaptation originale qui vise à n'adapter que les probabilités des n-grammes contenant des termes liés au thème et à laisser les autres probabilités inchangées. Concrêtement, notre processus est illustré par la figure 6.2 : à partir d'un corpus thématique, nous cherchons à construire automatiquement une terminologie, c'est-à-dire une liste de mots ou séquences de mots qui portent une notion du thème considéré, que nous utilisons ensuite pour contraindre la réestimation du modèle de langue généraliste grâce à une adaptation par minimum d'information discriminante (MDI) (Lecorvé *et al.*, 2009). Avant de détailler cette contribution, nous rappelons le fonctionnement général de l'adaptation MDI puis donnons une vision d'ensemble des travaux liés à l'extraction de terminologies.

6.2.1 Adaptation MDI

L'objectif de l'adaptation MDI est de trouver un modèle de langue dont la distribution de probabilités P_A satisfait un ensemble de contraintes linéaires et possède une entropie relative (ou divergence de Kullback-Leibler) minimale avec la distribution d'un modèle de langue de départ. Cette méthode est connue pour la souplesse et l'expressivité qu'elle offre *via* son mécanisme de construction de contraintes. Nous en donnons ici une description détaillée pour une tâche d'adaptation générique.

6.2.1.1 Principe général

Considérons une distribution de départ P_B sur un ensemble de n-grammes V^n, V étant le vocabulaire du système, ainsi qu'un ensemble de caractéristiques définies arbitrairement et permettant de rassembler des n-grammes au sein de différentes classes en fonction de propriétés qu'ils partagent. Par exemple, une première caractéristique peut consister à grouper les n-grammes contenant un mot donné, une autre à grouper ceux qui terminent par un mot au pluriel... De plus, chaque caractéristique i est associée à une masse de probabilité K_i reflétant son importance dans une source d'information observée, par exemple un corpus textuel. Le principe de l'adaptation MDI est alors de rechercher la distribution adaptée P_A qui résoud un système de contraintes linéaires construit à partir des caractéristiques, où chaque contrainte i restreint la masse K_i à être distribuée entre tous les n-grammes qui présentent la caractéristique i. Mathématiquement, ces contraintes s'écrivent :

$$\langle f_i, P_A \rangle = K_i, \quad \forall i \in [1..k] \tag{6.1}$$

$$\text{avec } \langle f_i, P_A \rangle = \sum_{hw \in V^n} f_i(hw) \times P_A[hw] \ , \tag{6.2}$$

où $f_i(hw)$ est une fonction indicatrice (*feature function*) de la caractéristique i, définie par :

$$f_i(hw) = \begin{cases} 1 & \text{si } hw \text{ présente la caractéristique } i, \\ 0 & \text{sinon .} \end{cases} \tag{6.3}$$

Parmi l'ensemble des solutions de (6.1), P_A est alors définie comme celle qui reste la plus proche de P_B, c'est-à-dire la distribution qui minimise sa divergence de Kullback-Leibler avec P_B :

$$P_A = \arg\min_P D_{KL}(P||P_B), \tag{6.4}$$

$$\text{avec } D_{KL}(P||P_B) = \sum_{hw \in V^n} P[hw] \log \frac{P[hw]}{P_B[hw]} \ . \tag{6.5}$$

En supposant que P_A est de la forme suivante :

$$P_A[hw] = \frac{P_B[hw]}{Z(hw)} \times \prod_{i=1}^{K} e^{\lambda_i \times f_i(h,w)} \tag{6.6}$$

$$= P_B[hw] \times \alpha(hw) \ , \tag{6.7}$$

où $Z(hw)$ est un facteur de normalisation, K est le nombre de caractéristiques et les paramètres λ_i sont des constantes à déterminer, il a été montré que le facteur de mise à

l'échelle $\alpha(hw)$ de chaque n-gramme hw peut être calculé de manière itérative grâce l'algorithme *Generalized Iterative Scaling* (GIS) (Darroch et Ratcliff, 1972). Après simplification des calculs de cet algorithme itératif, la probabilité jointe $P_A[hw]$ d'un n-gramme hw peut s'exprimer comme :

$$P_A[hw] = P_B[hw] \times \underbrace{\prod_{i=1}^{K} \left(\frac{K_i}{\langle f_i, P_B \rangle} \right)^{\frac{f_i(hw)}{\chi_{hw}}}}_{\alpha(hw)} , \tag{6.8}$$

où χ_{hw} est le nombre de caractéristiques auxquelles participe hw.

6.2.1.2 Cas d'un modèle *n*-gramme avec *backoff*

Dans le cadre d'un modèle de langue n-gramme avec *backoff*, la formule (6.8) conduit à calculer la probabilité conditionnelle adaptée d'un n-gramme comme :

$$P_A[w|h] = \frac{P_B[w|h] \times \alpha(hw)}{\sum_{\hat{w} \in V} P_B[\hat{w}|h] \times \alpha(h\hat{w})} = \frac{P_B[w|h] \times \alpha(hw)}{Z(h)} \tag{6.9}$$

et au coefficient de repli adapté de l'historique h :

$$\beta_A(h) = \beta_B(h) \times \frac{\alpha(hw)}{\alpha(h^- w)} \tag{6.10}$$

où $\beta_B(h)$ est le coefficient de repli dans le modèle de langue initial et h^- désigne l'historique h ramené à son ordre inférieur.

En pratique, comme le calcul du coefficient de normalisation $Z(h)$ peut être très lourd lorsque la taille du vocabulaire et l'ordre du modèle de langue deviennent grands, il a été proposé de contraindre en outre la masse de probabilité de tout historique h à être conservée durant l'adaptation (Kneser *et al.*, 1997), soit la contrainte :

$$mass(h) = \sum_{hw \in E} P_B[w|h] = \sum_{hw \in E} P_A[w|h] , \tag{6.11}$$

où E est l'ensemble des n-grammes ayant été observés lors de l'apprentissage de la distribution initiale P_B. Cette contrainte induit une nouvelle expression du coefficient de normalisation où ne sont considérés que les n-grammes ayant été observés lors de l'apprentissage du modèle de langue initial et non plus l'ensemble des n-grammes possible à partir du vocabulaire du système :

$$Z'(h) = \frac{\sum_{\boldsymbol{hw \in E}} P_B[w|h] \times \alpha(hw)}{\sum_{\boldsymbol{hw \in E}} P_B[w|h]} . \tag{6.12}$$

Ce cadre générique d'adaptation MDI est très précisément celui que nous avons développé et intégré au sein du système IRENE et de son décodeur SIROCCO.

6.2.1.3 Mise en pratique

Étant donné un corpus d'adaptation – dans notre cas, un corpus thématique – , l'adaptation MDI consiste par défaut à considérer chaque n-gramme possible à partir du vocabulaire V comme une caractéristique et à poser l'ensemble de contraintes :

$$\langle f_{hw}, P_A \rangle = \langle f_{hw}, P_a \rangle, \quad \forall hw \in V^n \ , \tag{6.13}$$

où P_a est une distribution apprise à partir du corpus d'adaptation.

Comme ce corpus est généralement de taille modeste et ne permet donc pas d'estimer de manière fiable des n-grammes trop longs, les n-grammes sont traditionnellement regroupés en fonction de leur seul dernier mot et la réestimation est contrainte par des probabilités unigrammes (Kneser *et al.*, 1997), soit :

$$\langle f_{\hat{w}}, P_A \rangle = P_a[\hat{w}], \quad \forall \hat{w} \in V \ , \tag{6.14}$$

où $f_{\hat{w}}$ est définie par :

$$f_{\hat{w}}(hw) = \begin{cases} 1 & \text{si } w = \hat{w}, \\ 0 & \text{sinon .} \end{cases} \tag{6.15}$$

Cette méthode, appelée *unigram rescaling*, conduit à un facteur de mise à l'échelle relativement simple :

$$\alpha(hw) = \alpha(w) = \left(\frac{P_a[w]}{P_B[w]} \right)^{\gamma} \ , \tag{6.16}$$

où γ est un facteur ajouté artificiellement pour lisser le rapport des probabilités unigrammes (Kneser *et al.*, 1997 ; Federico, 1999 ; Tam et Schultz, 2006). Ce facteur est généralement optimisé sur un ensemble de développement. Néanmoins, dans nos expériences, nous ne cherchons pas à optimiser cette valeur car nous nous focalisons sur d'autres aspects de l'adaptation MDI. Ainsi, la valeur γ sera toujours fixée à 1 au cours de nos expériences.

Maintenant que nous avons décrit le fonctionnement de l'adaptation MDI, nous présentons quelques notions relatives à la tâche d'extraction de terminologies.

6.2.2 Notions de terminologie

En tant que discipline, la *terminologie* s'intéresse à l'étude des notions et des termes en usage au sein d'un langage de spécialité. Il s'agit d'un domaine largement étudié tant en linguistique classique qu'en linguistique computationnelle. Nous ne prétendons pas ici en donner une description détaillée mais souhaitons aborder les quelques éléments-clés dont nous avons besoin pour la suite de ce chapitre.

Sur un plan linguistique, la construction d'une terminologie consiste à dresser une liste de termes faisant référence de façon univoque à des notions propres à un domaine donné. Alors que ces termes peuvent êtres donnés par un expert du domaine, nous nous intéressons à leur extraction automatique à partir d'un corpus textuel spécialisé. Parmi ces termes, une distinction est faite entre termes simples et termes complexes. Les premiers se composent d'un unique mot plein, par exemple un nom. Ces termes simples ont l'inconvénient d'être parfois ambigus mais présentent l'intérêt d'avoir un comportement syntaxique simple à modéliser, ce qui facilite leur extraction à partir d'un corpus et leur utilisation pour différentes tâches. Les termes complexes sont quant à eux constitués de plusieurs unités lexicales pleines et sont ainsi généralement nettement moins ambigus. Ils présentent

toutefois l'inconvénient de pouvoir se décliner en de multiples variantes. Par exemple, les séquences « service financier », « services des comptes et finances » et « service des finances » se rapportent tout trois à la même notion de « service qui est en charge de la gestion des finances » alors que, sur un plan syntaxique, ces séquences sont différentes. Bien que des typologies de ces variantes aient été proposées dans (Daille, 2002 ; Ibekwe-SanJuan et SanJuan, 2003), ce phénomène de variation terminologique complique considérablement la modélisation des termes complexes en vue de leur extraction en corpus. Différentes méthodes existent toutefois dans la littérature tant pour l'extraction de termes simples et que pour celle de termes complexes à partir de corpora spécialisés.

6.2.2.1 Termes simples

À notre connaissance, les travaux actuels d'extraction de termes simples à partir d'un corpus de spécialité se reposent essentiellement sur des comparaisons de fréquences d'apparitions des unités lexicales (quelle que soit leur nature) dans le corpus de spécialité par rapport à celles relevées dans un corpus généraliste (Drouin, 2004 ; Rossignol, 2005). D'après un seuil de significativité (Chung, 2003) ou d'après des tests statistiques (Kilgarriff, 2001), les unités apparaissant un certain nombre de fois plus souvent dans le corpus de spécialité que dans le corpus généraliste sont considérées comme des termes simples.

6.2.2.2 Termes complexes

Les travaux qui s'intéressent à l'extraction de termes complexes sont quant à eux plus variés. Ils s'appuient sur trois approches : l'une statistique, l'autre symbolique, la dernière mixte.

Les approches statistiques se fondent sur la recherche de collocations, c'est-à-dire la recherche d'unités lexicales qui apparaissent anormalement fréquemment ensemble (par exemple, *filet* et *mignon* si l'on parle de cuisine). Étant donné un corpus, la recherche de ces collocations s'effectue généralement par le dénombrement de co-occurrences de mots au sein d'une fenêtre glissante de taille fixe (Church et Hanks, 1990) ou d'arbres syntaxiques (Grefenstette et Teufel, 1995). La force du lien de chaque paire de mots est alors mesurée, ce qui permet d'ordonner l'ensemble des paires et d'exclure par un seuillage celles supposées inintéressantes. Différentes mesures sont utilisées dans la littérature comme, par exemple, l'information mutuelle (Church et Hanks, 1990), l'information mutuelle au cube (Daille, 1994), le coefficient de Dice (Jacquemin, 1997) ou encore le rapport de log-vraisemblance (Dunning, 1993).

De leur côté, les approches symboliques reposent sur le fait que les termes complexes suivent souvent un patron syntaxique stable. Par exemple en français, beaucoup de termes complexes sont de la forme `<Nom Préposition Verbe_à_l'infinitif>` tels que *« pince à épiler »*, *« machine à coudre »*... L'extraction des termes complexes consiste alors en la découverte de patrons syntaxiques propres aux termes complexes, puis en la recherche de séquences de mots correspondant à l'un de ces patrons (Justeson et Katz, 1995 ; Strzalkowski, 1995 ; Ibekwe-SanJuan, 2001). Dans les faits, les patrons syntaxiques peuvent être donnés par un expert (Roche *et al.*, 2004) ou appris automatiquement (Ramshaw et Marcus, 1995 ; Claveau et Sébillot, 2004).

Enfin, beaucoup de méthodes sont mixtes et tirent profit des intérêts des approches statistiques et symboliques par l'enchainement, parfois itératif, de chacune d'entre elles, que ce soit dans un ordre ou dans l'autre. L'idée de ces travaux est de dresser une liste de

termes-candidats à partir d'une première approche et de filtrer ces candidats par l'entremise d'une seconde approche (Smadja, 1993 ; Frantzi *et al.*, 2000 ; Daille, 2002 ; Claveau et Sébillot, 2004).

Dans notre travail, nous choisissons d'étudier l'utilité des termes simples et des termes complexes extraits à partir de corpora thématiques pour restreindre l'adaptation d'un modèle de langue aux seuls n-grammes qui contiennent des termes du thème. Nous apportons tout d'abord une attention particulière à l'utilisation de termes simples, puis nous étudierons celle de termes complexes.

6.3 Utilisation de termes simples

Notre idée fondamentale pour l'adaptation thématique d'un modèle de langue est que seules certaines probabilités de n-grammes liés au thème doivent être modifiées alors que l'ensemble des autres probabilités doit rester inchangé étant donné qu'elles portent sur des n-grammes n'ayant aucune spécificité au sein du thème et qu'elles doivent donc rester égales à celles apprises sur un corpus généraliste. Notre approche consiste alors à guider une adaptation MDI par une terminologie de termes simples extraits à partir de nos corpora thématiques, et à ne réestimer que les probabilités conditionnelles des n-grammes se terminant par l'un de ces termes. Par ailleurs, nous pensons également que certains mots jouent un rôle similaire au sein d'un thème. Aussi, nous proposons d'intégrer des connaissances linguistiques au sein du processus d'adaptation pour vérifier cette hypothèse. Dans cette section, nous présentons tout d'abord notre méthode d'extraction de termes simples avant de décrire différentes techniques de réestimation.

6.3.1 Extraction de termes simples

Notre approche d'extraction de termes simples s'appuie sur le critère *tf-idf*. Pour chaque document d d'un corpus thématique, nous calculons son vecteur de scores :

$$\overrightarrow{v_d} = \begin{pmatrix} S'_d(\ell_1) \\ \vdots \\ S'_d(\ell_N) \end{pmatrix} , \tag{6.17}$$

où $S'_d(\ell_i)$ est le score *tf-idf* du lemme ℓ_i dans le document d tel que présenté à la section 5.1 et N est le nombre de formes distinctes dans d. À partir de ces scores, nous calculons le vecteurs moyen $\overrightarrow{T}$ pour l'ensemble du corpus thématique $\mathcal{C}$:

$$\overrightarrow{T} = \frac{1}{|\mathcal{C}|} \times \sum_{d \in \mathcal{C}} \overrightarrow{v_d} , \tag{6.18}$$

où $|\mathcal{C}|$ est le nombre de documents dans $\mathcal{C}$. Au sein de ce vecteur, les lemmes ayant les scores les plus élevés sont ceux supposés être les plus représentatifs du thème. Après avoir remplacé chaque lemme par l'ensemble de ses flexions possibles[28], nous définissons alors une terminologie du corpus thématique $\mathcal{C}$ comme la liste des n mots dont les lemmes ont les meilleurs scores dans $\overrightarrow{T}$. Nous notons cette terminologie T_n. La table 6.4, qui donne la liste

[28] L'extension d'un lemme à l'ensemble de ses flexions possibles dans la langue est effectuée à partir d'une liste de correspondances entre lemmes et flexions relevées à partir de la lemmatisation des articles extraits des archives du journal *Le Monde* balayant la période 1987–2007.

laïcités	laïque	voilées	élèves	juives
laïcité	école	républiques	élève	juifs
musulmans	écoles	république	catholiques	juif
musulman	foulards	sarkozy	catholique	églises
musulmanes	foulard	france	question	église
musulmane	islamiques	publiques	questions	filles
voile	islamique	public	républicaines	fille
voiles	ports	publique	républicains	stasi
religieuses	port	publics	républicain	françaises
religieuse	culte	droits	républicaine	français
religieux	cultes	droit	discriminations	française
religion	liberté	débat	discrimination	communautés
religions	libertés	débats	hommes	communauté
islam	signe	paris	homme	jacques
islams	signes	mosquée	pays	croyances
femmes	principes	mosquées	dieu	croyance
femme	principe	chrétiennes	dieux	ramadan
loi	voilés	chrétien	séparation	ramadans
lois	voilé	chrétiens	séparations	scolaire
laïques	voilée	chrétienne	juive	scolaires

TABLE 6.4 – Liste des 100 mots ayant les meilleurs scores pour un corpus de 200 pages construit à partir d'un segment traitant de la laïcité et du port du voile.

des termes de T_{100} pour un segment traitant de la laïcité et du port du voile, permet de constater que cette méthode fournit de bons résultats dans l'ensemble, malgré la présence de quelques termes non-spécifiques au thème (Sarkozy, France, Paris, français, Jacques...). Ce léger bruit est lié au fait que ces mots sont très fréquents dans le corpus thématique bien qu'ils ne partagent pas de lien sémantique particulier avec le reste des autres termes[29]. Somme toute, la bonne qualité générale de ces listes de termes simples permet tout de même de mettre en œuvre une adaptation MDI spécifique à chaque thème.

6.3.2 Spécialisation de l'*unigram rescaling*

Travaillant à partir de termes simples, nous avons voulu expérimenter différentes stratégies d'adaptation en reprenant le cadre de l'*unigram rescaling*. Ces stratégies s'appuient sur deux hypothèses que nous souhaitons vérifier. D'une part, nous pensons qu'une adaptation thématique ne doit pas s'appuyer sur tous les mots du vocabulaire du système mais seulement sur les termes du thème. D'autre part, nous estimons que certains mots partagent un même rôle sémantique au sein d'un thème et que, par conséquent, des n-grammes contenant ces mots devraient voir leur probabilité être adaptée dans de mêmes proportions.

6.3.2.1 Restriction à l'ensemble des termes simples

Pour ne pas procéder à une adaptation de toutes les probabilités n-grammes, nous proposons d'aménager la méthode d'*unigram rescaling* en regroupant uniquement au sein de

[29] Suite à des réflexions postérieures aux travaux que nous présentons ici, nous montrons dans l'annexe D que ces terminologies peuvent être améliorées par le choix d'une autre variante du score *tf-idf* que celle utilisée ici et par l'utilisation de relations paradigmatiques.

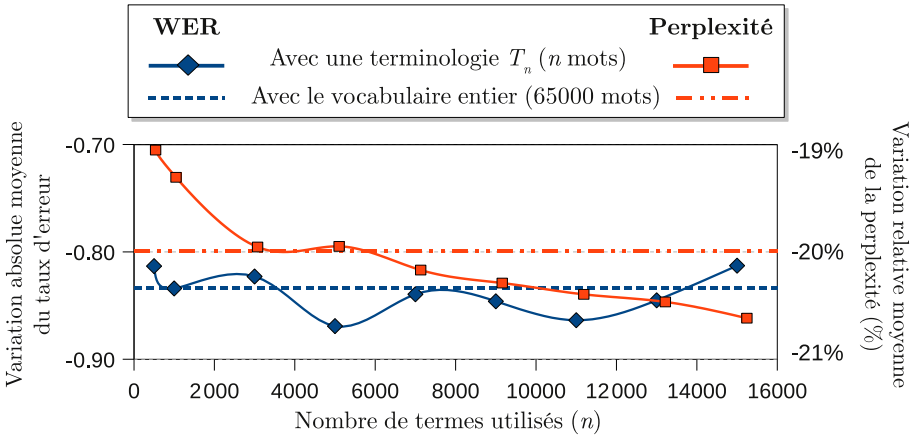


FIGURE 6.3 – Influence du nombre de termes considérés pour l'*unigram rescaling* sur le WER et la perplexité.

mêmes caractéristiques les n-grammes terminant par un même terme simple. Ceci s'exprime par l'ensemble de contraintes suivant :

$$\langle f_{\hat{w}}, P_A \rangle = P_a[\hat{w}], \quad \forall \hat{w} \in T_n \ , \tag{6.19}$$

où T_n est une terminologie de n mots apprise à partir d'un corpus thématique et P_a est une distribution unigramme estimée sur ce même corpus. Ce système conduit à un facteur de mise à l'échelle suivant pour un terme t :

$$\alpha(ht) = \alpha(t) = \frac{P_a[t]}{P_B[t]} \ . \tag{6.20}$$

Pour le reste des n-grammes ne terminant pas par un terme simple, il peut être démontré à partir de la formule (6.8) que cet ensemble de contraintes conduit à leur attribuer un facteur de mise à l'échelle égal à 1 (*cf.* l'annexe A.3), c'est-à-dire que les probabilités de ces n-grammes sont directement reportées à partir de la distribution du modèle généraliste, à un coefficient de normalisation près.

Nous avons comparé l'utilisation de terminologies de différentes tailles par rapport à la prise en compte du vocabulaire entier pour l'adaptation de notre modèle généraliste. La figure 6.3 montre l'impact de ces deux stratégies sur notre ensemble de développement en terme de WER et de perplexité. Il apparaît que, quelque soit leur taille, l'utilisation de terminologies conduit à peu près aux mêmes gains sur le WER et sur la perplexité que l'adaptation fondée sur le vocabulaire complet. Il semblerait qu'une taille de 5000 termes soit suffisante puisqu'il s'agit de la valeur aboutissant aux meilleurs gains sur le WER. Ces premiers résultats tendent à montrer que seuls les termes liés au thème contribuent efficacement à l'adaptation thématique.

Une deuxième série d'expériences a été menée en ne considérant cette fois que des terminologies de 500 termes (T_{500}), ce qui constitue un cas extrême puisque cela représente moins de 1 % du vocabulaire, et de 5000 termes (T_{5000}), valeur optimale pour le WER. Ces expériences ont consisté à mesurer l'effet de l'adaptation MDI lorsque celle-ci incorpore ou exclut les mots de nos terminologies au sein de l'ensemble des caractéristiques servant à l'adaptation MDI. Les résultats en terme de perplexité, WER et LER de ces expériences sont présentés et comparés à ceux obtenus *via* une interpolation linéaire par la table 6.5.

		Perplexité	WER	LER
Sans adaptation		96,9	22,1	19,4
Interpolation linéaire ($\lambda = 0,8$)		80,1 (−17 %)	21,4 (−0,7)	18,6 (−0,8)
Adaptation MDI fondée sur	V	**75,5** (−20 %)	21,3 (−0,8)	**18,3** (−1,1)
	T_{500}	76,7 (−19 %)	21,3 (−0,8)	18,5 (−0,9)
	T_{5000}	**75,5** (−20 %)	**21,2** (−0,9)	18,4 (−1,0)
	V − T_{500}	94,8 (−2 %)	21,9 (−0,2)	19,3 (−0,1)
	V − T_{5000}	95,4 (−2 %)	22,0 (−0,1)	19,4 (0,0)

TABLE 6.5 – Perplexité, WER et LER mesurés sur l'ensemble de développement sans adaptation et avec différentes méthodes d'adaptation. Entre parenthèses, les variations relatives moyennes de la perplexité et les variations absolues moyennes des taux d'erreur.

Outre les meilleurs résultats de l'adaptation MDI par rapport à l'interpolation linéaire, ces résultats montrent clairement que l'adaptation n'a pas d'effet lorsque les termes simples sont exclus de l'adaptation thématique (cas « V − T_{500} » et « V − T_{5000} ») , ce qui confirme notre hypothèse de travail.

6.3.2.2 Regroupement de termes simples

Notre seconde hypothèse pour rendre plus efficace une adaptation thématique est que certains mots partagent un même rôle sémantique dans le langage utilisé au sein d'un thème. Au delà du choix des termes à utiliser pour l'adaptation MDI, cette hypothèse implique de s'intéresser aux choix des fonctions caractéristiques utilisées pour regrouper les n-grammes du modèle initial. Au lieu de regrouper les n-grammes en fonction de leur dernier mot, comme cela se fait dans la technique d'*unigram rescaling*, nous proposons trois nouvelles fonctions caractéristiques : une première fondée sur des lemmes :

$$f_\ell(hw) = \begin{cases} 1 & \text{si } \ell \text{ est le lemme de } w, \\ 0 & \text{sinon ,} \end{cases} \tag{6.21}$$

une seconde qui, pour chaque terme $\hat{w}$, rassemble tous les n-grammes dont le dernier mot partage une relation paradigmatique[30] avec $\hat{w}$:

$$f_{\mathcal{R}(\hat{w})}(hw) = \begin{cases} 1 & \text{si } w \text{ partage une relation paradigmatique avec } \hat{w}, \\ 0 & \text{sinon ,} \end{cases} \tag{6.22}$$

et une dernière qui regroupe tous les mots d'une terminologie T donnée au sein d'une même caractéristique :

$$f_T(hw) = \begin{cases} 1 & \text{si } w \text{ appartient à } T, \\ 0 & \text{sinon .} \end{cases} \tag{6.23}$$

La fonction caractéristique f_ℓ se justifie par l'idée selon laquelle l'information flexionnelle n'est pas importante pour juger de l'importance d'un mot au sein d'un thème, c'est-à-dire que différentes flexions d'un même terme partagent la même importance. Si cette idée peut se discuter dans certains cas, les termes liberté et libert**és** ne réferrant pas forcément le

[30] Ces relations sont issues d'un apprentissage automatique.

		Perplexité	WER	LER
Sans adaptation		96,9	22,1	19,4
T_{500}	Pas de regroupement ($f_{\hat{w}}$)	76,7 (−19%)	21,3 (−0,8)	18,5 (−0,9)
	Regroupement par lemme (f_{ℓ})	77,0 (−19%)	21,4 (−0,7)	18,7 (−0,7)
	Regroupement paradigmatique ($f_{\mathcal{R}(\hat{w})}$)	91,6 (−5%)	21,8 (−0,3)	19,1 (−0,3)
	Tous regroupés (f_T)	89,2 (−7%)	21,7 (−0,4)	19,1 (−0,3)
T_{5000}	Pas de regroupement ($f_{\hat{w}}$)	75,5 (−20%)	21,2 (−0,9)	18,4 (−1,0)
	Regroupement par lemme (f_{ℓ})	74,2 (−20%)	21,4 (−0,7)	18,6 (−0,8)
	Regroupement paradigmatique ($f_{\mathcal{R}(\hat{w})}$)	90,0 (−6%)	21,8 (−0,3)	19,0 (−0,4)
	Tous regroupés (f_T)	94,0 (−3%)	21,9 (−0,2)	19,3 (−0,1)

TABLE 6.6 – Perplexité, WER et LER mesurés sur l'ensemble de développement pour différentes fonctions caractéristiques pour des terminologies de 500 et 5000 termes.

même concept dans l'absolu, elle est globalement défendable, d'autant plus quand on se place dans un domaine donné. La fonction $f_{\mathcal{R}(\hat{w})}$ se fonde sur le fait que l'existence du lien paradigmatique entre deux mots signifie que ces deux mots apparaissent dans des contextes lexicaux similaires. À partir du moment où bicyclette est un terme, cette fonction vise, par exemple, à adapter dans les mêmes proportions les n-grammes « monte à bicyclette », « monte à vélo », « vend des vélo » ou encore « arrête de pédaler ». La fonction f_T est, quant à elle, plus catégorique puisqu'elle repose sur l'hypothèse selon laquelle tous les mots d'une terminologie participent de la même manière à l'évocation d'un thème.

La table 6.6 présente les résultats obtenus en termes de perplexité, WER et LER à partir de ces différentes fonctions caractéristiques et les comparent à ceux obtenus avec la fonction caractéristique classique $f_{\hat{w}}$. Il apparaît que le regroupement de tous les termes simples au sein d'une même caractéristique (fonction f_T) produit des gains nettement plus faibles que ceux reportés pour les autres fonctions. Ces résultats sont d'ailleurs d'autant moins bons que le nombre de termes considérés augmente. Nous expliquons ces résultats par le fait que notre hypothèse sur l'importance équivalente de tous les termes au sein d'un thème est probablement trop grossière. Un constat similaire peut être fait en terme de résultats pour l'utilisation de relations paradigmatiques. Nous pensons toutefois que l'hypothèse sur laquelle s'appuie cette idée reste bonne. Nous attribuons plutôt ces résultats en demi-teintes à la qualité médiocre de nos relations paradigmatiques utilisées dans ces expériences, ce qui tend à réduire la spécificité de chaque caractéristique et conduit à adapter un peu tous les n-grammes du modèle de langue dans des proportions relativement similaires. Il serait intéressant de réitérer ces expériences avec des relations de meilleure qualité. De son côté, la stratégie de regroupement en lemmes produit des résultats quasi similaires à ceux obtenus avec la fonction habituelle $f_{\hat{w}}$. Plus précisément, les gains en perplexité sont identiques alors que les WER et LER retournés sont légèrement moins bons. À la lumière de ces résultats, nous avons choisi d'écarter les fonctions caractéristiques f_T et $f_{\mathcal{R}(\hat{w})}$ pour nos expériences sur le corpus de test.

6.3.3 Résultats sur le corpus de test

Suite à nos expériences sur notre ensemble de développement, nous avons cherché à vérifier les différents résultats obtenus sur notre ensemble de test. Ces résultats sont présentés par la table 6.7 en terme de perplexité, WER et LER.

		Perplexité	WER	LER
Sans adaptation		96,7	20,7	18,3
Interpolation linéaire ($\lambda = 0,8$)		78,8 (−16%)	20,2 (−0,5)	17,4 (−0,9)
MDI basé sur	$f_{\hat{w}}$ + V	**74,7 (−21%)**	20,1 (−0,6)	17,3 (−1,0)
	$f_{\hat{w}}$ + T_{500}	76,5 (−19%)	20,2 (−0,5)	17,4 (−0,9)
	$f_{\hat{w}}$ + T_{5000}	75,5 (−20%)	**20,0 (−0,7)**	**17,2 (−1,1)**
	f_{ℓ} + T_{500}	76,8 (−19%)	20,2 (−0,5)	17,5 (−0,8)
	f_{ℓ} + T_{5000}	75,7 (−20%)	20,1 (−0,6)	17,2 (−1,1)

TABLE 6.7 – Perplexité, WER et LER mesurés sur l'ensemble de test sans adaptation et avec différentes méthodes d'adaptation. Entre parenthèses, les variations relatives moyennes de la perplexité et les variations absolues moyennes des taux d'erreur.

STUDENT	Inter. lin.	$f_{\hat{w}}$ + V	$f_{\hat{w}}$ + T_{500}	$f_{\hat{w}}$ + T_{5000}
Sans adaptation	$7,7 \times 10^{-5}$	$8,8 \times 10^{-5}$	$3,4 \times 10^{-7}$	$\mathbf{1,5 \times 10^{-7}}$
Interpolation linéaire	–	0,44	0,67	0,22
$f_{\hat{w}}$ + V	–	–	0,67	0,63
WILCOXON	**Inter. lin.**	$f_{\hat{w}}$ + V	$f_{\hat{w}}$ + T_{500}	$f_{\hat{w}}$ + T_{5000}
Sans adaptation	$1,3 \times 10^{-4}$	$2,5 \times 10^{-4}$	$1,2 \times 10^{-6}$	$\mathbf{3,3 \times 10^{-7}}$
Interpolation linéaire	–	0,83	0,87	0,26
$f_{\hat{w}}$ + V	–	–	0,77	0,45

TABLE 6.8 – Significativités statistiques d'après les tests de Student et de Wilcoxon entre les WER mesurés sans adaptation et avec différentes stratégies d'adaptation MDI. Le taux de confiance de ces tests est fixé à 0,05.

Tout d'abord, de même que pour le résultats reportés à la section 6.1, nous pouvons remarquer que les résultats par défaut sur le corpus de test sont nettement meilleurs que ceux obtenus sur le corpus de développement. Plus précisément, un écart absolu de 1,4 est reporté sur le WER par défaut, c'est-à-dire obtenu sans adaptation, par rapport à celui précédemment reporté pour l'ensemble de développement (tables 6.5 et 6.6). Ainsi, les gains reportés pour l'ensemble des techniques expérimentées sont globalement moindres.

Ensuite, il apparaît encore une fois que l'utilisation de terminologies conduit à des gains comparables à ceux obtenus lorsque des contraintes sont portées sur l'ensemble des n-grammes du modèle de langue initial et que ces gains sont meilleurs que ceux obtenus par une interpolation linéaire. Comme le montre les valeurs de la table 6.8, ces faibles variations observées sur le WER d'une méthode à l'autre ne sont toutefois pas statistiquement significatives mais on remarque que l'utilisation de terminologies produit des gains beaucoup plus significatifs par rapport aux résultats du système généraliste, notamment dans le cas d'une terminologie de 5 000 termes.

Finalement, on remarque que l'utilisation de fonctions caractéristiques basées sur des regroupements en lemmes produits encore une fois des gains légèrement moins bons que la fonction classique $f_{\hat{w}}$. De la même manière que pour la stratégie basée sur la fonction f_T, il est possible d'en déduire que notre hypothèse concernant le rôle de l'information flexionnelle pour décrire la langue au sein d'un thème est fausse ou tout du moins abusive. Nous pensons toutefois que ce manque d'efficacité vient plutôt de l'estimation des probabilités à partir du corpus thématique. En effet, pour un lemme donné ℓ, la masse de probabilité à répartir

entre tous les n-grammes se terminant par une flexion de ℓ se calcule comme :

$$\langle f_\ell, P_a \rangle = \sum_{\substack{hw \in V^n \\ w \to \ell}} P_a[w] \ , \tag{6.24}$$

où $w \to \ell$ signifie que w a pour lemme ℓ. Or, dans de nombreux cas, beaucoup de flexions possibles de ℓ sont peu fréquentes voire absentes dans le corpus thématique, ce qui conduit à une sous-estimation de la masse de probabilité du lemme d'un terme simple. Par voie de conséquence, cette sous-estimation produit des facteurs de mise à l'échelle moindre et donc une adaptation thématique moins marquée. Plus généralement, nous pensons que ce phénomène intervient dès lors qu'une caractéristique regroupe beaucoup de mots, comme, par exemple, pour les fonctions f_T et $f_{\mathcal{R}(\hat{w})}$, précédemment présentées. Différentes solutions pourraient viser à pallier ce problème. Il serait, par exemple, envisageable de calculer les probabilités P_a directement sur une version lemmatisée du corpus thématique ou de recourir à une technique d'analyse sémantique latente pour le calcul des probabilités de chaque terme, bien que nous ayons délibérément choisi de ne pas nous engager sur cette voie. Cette dernière solution ouvre la voie à une perspective plus générale qui consisterait à passer d'un cadre d'estimation probabiliste de chaque caractéristique de l'adaptation MDI à un cadre possibiliste, qui vise à évaluer l'incertitude qui entoure des évènements à partir de l'observation d'une source d'information éparse[31] (Dubois, 2006 ; Oger *et al.*, 2010). La masse de probabilité d'une caractéristique serait alors remplacée par sa mesure de possibilité. Dans un autre ordre d'idée, le remplacement des fonctions caractéristiques par des fonctions d'appartenance non binaires pourrait permettre d'accorder plus d'importance à certaines flexions dans le calcul de la masse de probabilité d'un lemme.

Dans l'ensemble, nos résultats montrent essentiellement que peu de termes simples suffisent pour effectuer une adaptation thématique. Nous notons notamment que la considération de 500 termes seuls aboutit déjà, pour un coût calculatoire par ailleurs nettement inférieur, à des gains sur le WER comparables à ceux de l'interpolation linéaire ou de l'*unigram rescaling* classique. Il ressort tout de même de nos expériences que le regroupement des termes au sein de mêmes caractéristiques ne semble pas produire d'effet positif supplémentaire. Alors que certains des critères de regroupement que nous avons expérimentés peuvent *a posteriori* sembler manquer de pertinence, nous pensons que l'utilisation de relations paradigmatiques reste une idée à approfondir, notamment en produisant des relations paradigmatiques de meilleures qualité. Une autre piste que nous avons exploré consiste à tirer également parti de termes complexes en sus des termes simples.

6.4 Utilisation de termes complexes

Les termes complexes permettent de lever des ambiguités d'interprétation inhérentes à certains termes simples. Par exemple, les termes port et voile peuvent faire référence au domaine du nautisme tout comme à celui de la religion alors que le terme complexe

[31] Pour un évènement donné, une possibilité égale à 0 signifie que l'evènement n'est pas possible alors qu'une possibilité à 1 signifie simplement que la réalisation de l'évènement ne serait pas surprenante, mais juste normale. Ceci distingue une possibilité d'une probabilité pour laquelle la valeur 1 signifie qu'un évènement sera systèmatique en comparaison des autres évènements. Le cadre possibiliste n'entretient pas ce phénomène de concurrence entre évènements dans la mesure où aucune contrainte de somme à 1 des possibilités d'un ensemble d'évènements n'est considérée.

« port du voile » ne laisse aucun doute possible quant au thème dont il est question. Par ailleurs, la recherche de termes complexes peut faire émerger des séquences de mots propres à un thème sans toutefois qu'aucun de ces mots ne puisse être considéré comme un terme simple – on dit de ces séquences particulières qu'elles ont une opacité sémantique[32].

Cette section vise donc à étudier si, à travers ces propriétés, l'utilisation de termes complexes peut conduire à une adaptation du modèle de langue de meilleure qualité que celle que nous obtenons en ne considérant que des termes simples. Nous présentons tout d'abord le procédé d'extraction de termes complexes que nous avons utilisé, puis nous exposons nos différentes expériences concernant leur intégration au sein d'une adaptation thématique fondée sur MDI.

6.4.1 Extraction de termes complexes

Comme nous l'avons vu en 6.2.2, l'extraction en corpus de termes complexes est une tâche ardue, notamment en raison du grand nombre de variantes linguistiques sous lesquelles peut se présenter une même notion du thème. Nous avons choisi de déléguer cette tâche au logiciel ACABIT (Daille, 1994). Ce logiciel s'appuie sur une première phase de repérage et de regroupement de termes-candidats présentant des patrons syntaxiques donnés puis sur une étape de filtrage de ces candidats d'après des critères statistiques. ACABIT produit alors en sortie une liste de classes de termes complexes nominaux auxquels sont associés des scores reflétant leur importance dans le corpus traité. Comme le montre un exemple de ces classes dans le tableau 6.9, ce logiciel gère bien le phénomène de variation et, malgré le bruit émanant de nos corpora thématiques, les premiers termes complexes renvoyés, c'est-à-dire ceux auxquels ACABIT associe les scores les plus élevés, sont globalement pertinents pour le thème considéré. Dans cette mesure, il est envisageable d'utiliser ses résultats pour adapter thématiquement notre modèle de langue généraliste.

6.4.2 Utilisation dans MDI

De la même manière que pour les termes simples, différentes stratégies ont été expérimentées pour l'intégration des termes complexes. Nous avons cherché à savoir si ces termes sont à privilégier par rapport aux termes simples, s'ils les complètent ou si, au contraire, ils n'apportent rien. Pour cela, étant donné un ensemble X de termes, nous avons testé deux stratégies d'utilisation des termes complexes : l'une où aucun regroupement n'est fait entre les termes t de X :

$$f_X^t(hw) = \begin{cases} 1 & \text{si } hw \text{ termine par le terme } t \in X, \\ 0 & \text{sinon ,} \end{cases} \tag{6.25}$$

et l'autre où les variantes d'un terme complexe sont regroupées au sein d'une même caractéristique τ :

$$f_X^\tau(hw) = \begin{cases} 1 & \text{si } hw \text{ termine par un terme de } \tau \subset X, \\ 0 & \text{sinon .} \end{cases} \tag{6.26}$$

Ces deux stratégies ont été expérimentées sur différents ensembles X. Dans un premier cas, X correspond à l'ensemble T_c des termes complexes fournis par ACABIT à partir d'un

[32] Une séquence de mots présente une *opacité sémantique* lorsque le sens de cette séquence diffère de celui résultant de la composition des sens de chaque mot qui la compose. Par exemple, l'expression « casser sa pipe » n'a rien à voir avec le fait de détruire un instrument servant à fumer.

signes visibles religieux signes religieux signe religieux signe culturel ou religieux signes ostensibles religieux signes distinctifs religieux signes de religion signe ostentatoire de religion signes distinctifs de religion signe ostentatoire de la religion signes visibles religieux signes religieux signe religieux signe culturel ou religieux signes ostensibles religieux signes distinctifs religieux insignes religieux	droits de l' homme droit de l' homme droit des hommes droits entre hommes
port volontaire du voile port de voile port du voile	droit de vote droits de vote
principe républicain de laïcité principe de laïcité principes fondamentaux de la laïcité principes républicains de laïcité principes de la république et de la laïcité principes de la laïcité principe de la laïcité principes de laïcité principe constitutionnel de laïcité	musulmans en france musulmans de france musulman en france musulmans turcs de france musulmans pratiquants de france musulmans laïcs de france musulman de la france musulmans laïques de france
nicolas sarkozy	président sortant de la république président de la république présidents de la république
	femmes juives et non musulmanes femmes musulmans femmes musulmanes femmes d' origine musulmane femme non musulmane femme musulmane femmes non pas musulmanes femmes juives non musulmanes femmes non musulmanes
	services publics service public les services et les lieux publics services publiques

TABLE 6.9 – Exemple de premiers 50 termes complexes fournis par ACABIT à partir d'un corpus traitant du port du voile et le la laïcité.

corpus thématique, soit, en moyenne, 57 000 variantes de termes complexes représentant 52 000 formes normales différentes[33]. Dans un second cas, cet ensemble est restreint aux seuls termes complexes qui contiennent un terme simple issu d'une extraction préalable de 5000 termes simples, ce qui ramène le nombre moyen de variantes de termes complexes par segment thématique à environ 44 000. Nous notons cet ensemble de termes complexes $X = T_c \cap T_s$ où T_s désigne une terminologie de termes simples. Enfin, dans un troisième cas, nous avons considéré l'union des termes simples et de tous les termes complexes. Nous notons alors $X = T_c \cup T_s$.

La table 6.10 présente les résultats obtenus pour les 6 configurations possibles à partir de ces paramètres en terme de perplexité et de WER sur notre ensemble de développement. Ces résultats sont comparés à ceux obtenus à partir de 5000 termes simples seuls sans aucun regroupement particulier (fonction $f_{\hat{w}}$). Principalement, il apparaît alors que, quelle que soit l'utilisation faite des termes complexes, ceux-ci ne permettent jamais de produire une meilleure adaptation que celle fondée sur la seule utilisation de termes simples. Dans le détail, nous constatons que les termes complexes utilisés seuls (fonctions $f^*_{T_c}$) ne produisent

[33]Le faible écart entre ces deux nombre s'explique par le fait des variantes de termes ne sont généralement observées que pour les seuls termes complexes les plus fréquents dans le corpus. Les autres termes complexes se présentent alors généralement sous une forme unique.

	Termes simples	Termes complexes					
	$T_{5000} + f_{\hat{w}}$	$f^t_{T_c}$	$f^t_{T_c \cap T_s}$	$f^t_{T_c \cup T_s}$	$f^\tau_{T_c}$	$f^\tau_{T_c \cap T_s}$	$f^\tau_{T_c \cup T_s}$
Perplexité	**75,5** (−20 %)	91,1 (−5 %)	84,5 (−12 %)	78,4 (−17 %)	90,7 (−6 %)	80,5 (−15 %)	77,5 (−18 %)
WER	**21,2** (−0,9)	22,0 (−0,1)	21,8 (−0,3)	21,5 (−0,6)	22,0 (−0,1)	21,6 (−0,5)	21,4 (−0,7)

TABLE 6.10 – Perplexité et WER mesurés sur l'ensemble de développement sans adaptation (généraliste) et avec différentes stratégies d'utilisation des termes complexes au sein de l'adaptation MDI. Entre parenthèses, les variations relatives moyennes de la perplexité et les variations absolues moyennes des taux d'erreur.

quasi aucun effet que ce soit en terme de perplexité ou de WER. Nous expliquons ces résultats décevants par le fait que les terminologies renvoyées par ACABIT recensent énormément de termes complexes et que certains d'entre eux, notamment ceux qui présentent les scores les plus faibles, sont décorrélés du thème. Nous devrions probablement restreindre les terminologies complexes à un nombre maximum de termes. Les résultats obtenus en filtrant les termes complexes par des termes simples vont d'ailleurs dans ce sens puisque de meilleurs résultats sont reportés dans ce cas (fonctions $f^*_{T_c \cap T_s}$). Assez logiquement, nous nous apercevons que ces résultats sont encore améliorés dès lors que les termes simples sont introduits en complément des termes complexes (fonctions $f^*_{T_c \cup T_s}$). Malgré tout, cette dernière configuration reste moins performante que la seule utilisation de termes simples.

Ces différentes expériences montrent donc dans l'ensemble que, contrairement à notre hypothèse, les termes complexes en tant que séquences de mots n'apportent rien à la caractérisation linguistique d'un thème offerte par les termes simples et, au contraire, qu'ils la dégradent. Ce résultat peut paraître décevant mais il nous apparaît en partie logique. Tout d'abord, bien que le logiciel ACABIT renvoie globalement des termes complexes de qualité, ces derniers sont uniquement de type nominal. Ainsi, des expressions comme « porter le voile » et notamment ses variantes flexionnelles « porte le voile », « portent le voile » ou « porte un voile » ne sont pas prises en compte lors de l'adaptation alors qu'elles l'étaient implicitement lors de l'utilisation du terme simple voile. Le choix d'un autre logiciel retournant des termes complexes d'autres types pourrait être expérimenté (Cabré Castellví *et al.*, 2003). Ensuite, bien que le logiciel ACABIT soit capable de regrouper les variantes d'un terme complexe, nos terminologies complexes restent toujours limitées par le contenu des corpora thématiques. Or, il est évident que ces corpora ne pourront jamais contenir l'ensemble des variantes possibles. Une solution serait d'être capable de générer celles-ci automatiquement. Malheureusement, les différentes typologies de variation proposées dans la littérature ne permettent pas ce genre d'opération car les différentes règles qu'elles rassemblent ne sont pas systématiques, c'est-à-dire que certains termes peuvent tolérer certains types de variation alors que d'autres termes ne le peuvent pas. Enfin, beaucoup de termes complexes ne correspondent pas à des n-grammes recensés par le modèle de langue généraliste car ceux-ci n'ont pas été rencontrés ou ont été élagués lors de son apprentissage. Ainsi, bien qu'une adaptation soit souhaitée pour certains termes, rien n'est fait en pratique car les n-gramme correspondant à ces termes ne sont pas ajoutés dans le modèle de langue par l'adaptation MDI. C'est un inconvénient à souligner concernant cette technique par rapport à l'utilisation d'une interpolation linéaire. Pour dépasser cette importante limite, une solution pourrait être de repartir du modèle généraliste non élagué et d'effectuer un élagage

en fonction du thème ciblé. Cette solution n'est toutefois pas complètement satisfaisante car l'absence de certains n-grammes dans le modèle généraliste n'est pas forcément due à l'élagage mais simplement au fait qu'ils n'ont jamais été observés précédemment. Il faudrait alors plutôt réfléchir à une technique d'ajout de n-grammes, problème que nous traitons au chapitre 7.

Finalement, il ressort de l'ensemble de nos expériences visant l'utilisation d'une terminologie pour effectuer une adaptation thématique que la solution la plus performante est une technique d'*unigram rescaling* fondée sur quelques centaines ou milliers de termes simples sans chercher à regrouper ceux-ci *via* des propriétés linguistiques. Pourtant, malgré une diminuation significative du WER, nous avons observé que notre méthode ne parvenait pas toujours à corriger certains termes simples initialement mal transcrits par le modèle généraliste. La prochaine section s'intéresse à ce problème en donnant quelques éléments de diagnostics quant aux freins et faiblesses liés à notre technique pour obtenir de gains supplémentaires sur le WER, notamment sur mots spécifiques au thème.

6.5 Éléments de diagnostic

Durant le processus de transcription d'un groupe de souffle, l'utilisation d'un modèle de langue adapté thématiquement vise à favoriser l'émergence d'hypothèses comportant des séquences probables au sein du thème considéré, séquences jusqu'alors sous-estimées par le modèle de langue généraliste. Pour ce faire, nous avons montré qu'il était préférable de se fonder sur un ensemble restreint de termes simples et d'intégrer les modèles de langue adaptés au plus tôt dans le processus de transcription, c'est-à-dire dès la création des graphes de mots. Notre technique permet alors d'obtenir une diminution significative du WER global initial du système de reconnaissance. L'observation des sorties du système nous conduit à expliquer ce résultat par la correction d'erreurs de transcription faites sur des termes simples porteurs de l'adaptation. En détails, on remarque toutefois que ces corrections sont fréquentes lorsque les termes initialement mal transcrits sont déjà présents dans les graphes de mots générés par le modèle de langue généraliste alors que l'impact de l'adaptation thématique n'est globalement pas suffisant pour insérer des hypothèses comportant de nouveaux termes au moment de la création des graphes de mots.

Dans cette section, nous mettons en avant deux explications à ce phénomène. L'une relève du processus de transcription dans le sens où nous montrons que les termes d'un thème font particulièrement les frais de l'élagage de l'espace de recherche utilisé lors de l'étape de création des graphes de mots – ceci en dépit de l'importance linguistique accrue que leur apporte un modèle de langue adapté. L'autre est liée à notre méthode d'adaptation qui calcule la masse de probabilité de chacune des caractéristiques de l'adaptation MDI à partir d'un corpus thématique souffrant d'éparsité des données. Nous soulignons alors le fait que, malgré l'adaptation du modèle de langue, certains n-grammes se terminant par un terme ont toujours une probabilité trop faible pour que les hypothèses de phrase qui les contiennent survivent à l'élagage linguistique ayant lieu lors de la création des graphes. À l'issue de l'étude de ces deux limites, nous concluons qu'il est difficile de concilier parfaitement dans un même unique modèle de langue les informations relatives au langage général avec celles relatives au langage utilisé au sein d'un thème particulier. Par conséquent, il s'avère compliqué de transcrire sur la base d'un seul modèle de langue tous les groupes de souffles, qu'ils soient marqués thématiquement ou non. Nous envisageons alors la possibilité

de fusionner les résultats fournis par un modèle généraliste avec ceux qui produit un modèle de langue thématiquement adapté. Précisons que, dans cette section, nous cherchons uniquement à donner des éléments de diagnostic sur ces différents problèmes et à ouvrir différentes pistes de travail. Ainsi, l'ensemble des expériences que nous rapportons ici est effectué sur le seul ensemble de développement.

6.5.1 Favoriser le modèle de langue

Comme la potentialité d'apparition d'un mot dans le résultat final d'un décodage est avant tout conditionnée par sa présence dans les graphes de mots, l'étape de création des premiers graphes est particulièrement importante. Pour rappel, cette étape est effectuée par un algorithme de recherche en faisceau qui s'appuie sur une politique d'élagage de l'espace de recherche permettant de réduire le temps de calcul en écartant les hypothèses en cours d'analyse qui ne satisfont pas certains critères (*cf.* section 1.2.3). Pour cela, à chaque trame du signal à décoder, chaque hypothèse en cours d'analyse se voit attribuer un score Q_h au sein duquel interviennent les vraisemblances fournies par le modèle acoustique et les probabilités du modèle de langue mises à l'échelle par un facteur exponentiel ψ. Sont alors élaguées toutes les hypothèses trop éloignées acoustiquement et linguistiquement de la meilleure hypothèse courante respectivement d'après un seuil acoustique δ_{AC} et un seuil linguistique δ_{ML}. Par ailleurs, à chaque trame, seules les M meilleures hypothèses actives sont conservées.

Dans cette logique, le calcul des scores Q_h repose toujours sur le même compromis entre scores acoustiques et scores linguistiques, compromis régi par le facteur ψ. Ainsi, des hypothèses précédemment écartées lors de la création des graphes de mots à partir du modèle de langue généraliste peuvent très bien être à nouveau écartées lors de l'utilisation des modèles adaptés, en dépit de l'importance accrue que peuvent donner ces derniers modèles aux n-grammes contenant des termes simples. À ce phénomène peuvent s'ajouter des conditions acoustiques difficiles, par exemple une mauvaise phonétisation, une prononciation erronée ou encore un locuteur parlant avec un accent régional ou étranger.

Pour pallier ce problème, nous proposons de rendre le seuillage acoustique plus tolérant et de donner plus d'importance au modèle de langue dans le calcul des scores Q_h lors de l'utilisation d'un modèle thématiquement adapté. En pratique, ceci consiste à diminuer δ_{AC} et à augmenter ψ. Pour contrebalancer l'augmentation du nombre d'hypothèses actives engendrée par la diminution du seuil acoustique δ_{AC}, le nombre M d'hypothèses de fin de mot conservées pour chaque trame est abaissé de manière à ce que le temps global de calcul reste le même par rapport à celui relatif aux réglages d'origine.

Le tableau 6.11 présente, à partir du corpus de développement, les taux d'erreur sur les mots dans les graphes de mots à l'issue de leur création (GER) et dans les transcriptions finales (WER) en s'appuyant soit sur le modèle de langue généraliste soit sur des modèles de langue adaptés d'après une terminologie de 500 termes simples (configuration $f_{\hat{w}}+T_{500}$). En outre, nous calculons les taux d'erreurs sur ces seuls termes simples dans le graphes de mots et dans les transcriptions finales, nouveaux scores que nous nommons respectivement GTER[34] et TER[35]. Tous ces taux sont mesurés pour les réglages d'origine du système et pour nos réglages modifiés. Tout d'abord, il ressort clairement que les réglages donnant plus d'importance au modèle de langue produisent des taux d'erreur nettement meilleurs.

[34]Pour *Graph Term Error Rate*.
[35]Pour *Term Error Rate*.

	Modèle de langue	GER	GTER	WER	TER
Réglages d'origine	Généraliste	8,9	5,6	21,8	14,9
	Adapté	8,6	3,8	21,0	12,0
Réglages modifiés	Généraliste	8,5	5,5	21,0	13,7
	Adapté	**8,1**	**3,6**	**20,5**	**11,3**

TABLE 6.11 – Taux d'erreurs sur les mots et les termes dans les graphes de mots (respectivement GER et GTER) et dans les transcriptions (respectivement WER et TER) mesurés sur l'ensemble de développement pour les réglages d'origine du décodage et pour ceux modifiés accordant une plus grande importance au modèle de langue.

Référence		cas probable de la maladie
Réglages d'origine	Modèle généraliste	cas probable de la MÊLÉES
	Modèle adapté	cas probable de la MALAISIE
Réglages modifiés	Modèle généraliste	cas probable de la MÊLÉES
	Modèle adapté	cas probable de la maladie

TABLE 6.12 – Exemple d'alignement d'un groupe de souffle de référence pour les réglages d'origine et modifiés, avec un ML généraliste ou adapté.

Ce premier constat indiquent en premier lieu que les réglages d'origine du système IRENE n'étaient pas optimaux[36]. Cette amélioration des résultats s'explique principalement par le fait que, à temps de décodage égal, la configuration modifiée introduit en moyenne 18 % de nouveaux mots dans les graphes de mots par rapport aux réglages d'origine. Cette richesse lexicale accrue permet notamment aux modèles de langue adaptés d'influencer plus fortement le choix des hypothèses de sortie. On remarque ainsi que le taux d'erreur sur les termes atteint sa plus basse valeur dans le cas où un modèle adapté est utilisé avec les réglages modifiés (dernière ligne). Ensuite, il apparaît que les gains initiaux impliqués par l'adaptation thématique (ligne 2) se cumulent en partie mais pas entièrement avec ceux obtenus par nos nouveaux réglages sans adaptation (ligne 3). Ceci s'explique par le fait que les nouveaux réglages permettent parfois d'eux-mêmes de corriger des erreurs sur des termes. Toutefois, les gains reportés pour l'utilisation conjointe des nouveaux réglages et de l'adaptation thématique (colonne 4) sont le signe d'une certaine complémentarité entre ces deux mécanismes. Le tableau 6.12 illustre ce propos en présentant des alignements d'un groupe de souffle extrait d'un document parlant de la pneumonie atypique où le locuteur prononce « *malédie* » au lieu de « *maladie* ». Alors que « *maladie* » n'apparaît dans aucun des trois premiers cas dans la transcription finale, soit à cause du problème acoustique, soit à cause d'une probabilité linguistique trop faible, la combinaison de nos nouveaux réglages et de l'adaptation thématique permet d'obtenir la bonne sortie. Malgré tout, les taux d'erreur mesurés sur les termes dans les graphes montrent que le fait d'accorder plus d'importance aux scores linguistiques lors de la création des graphes n'est pas mis à profit par les modèles adaptés. On remarque ainsi seulement un écart absolu de $0,2$ entre les GTER obtenus des modèles adaptés avec les anciens réglages et ceux obtenus avec les

[36] Ce constat n'a rien de surprenant étant donné le nombre important de paramètres co-existant au sein d'un système de reconnaissance automatique de la parole. Ainsi, la configuration d'origine retenue pour notre système n'avait pas fait l'objet d'une optimisation conjointe de tous les paramètres.

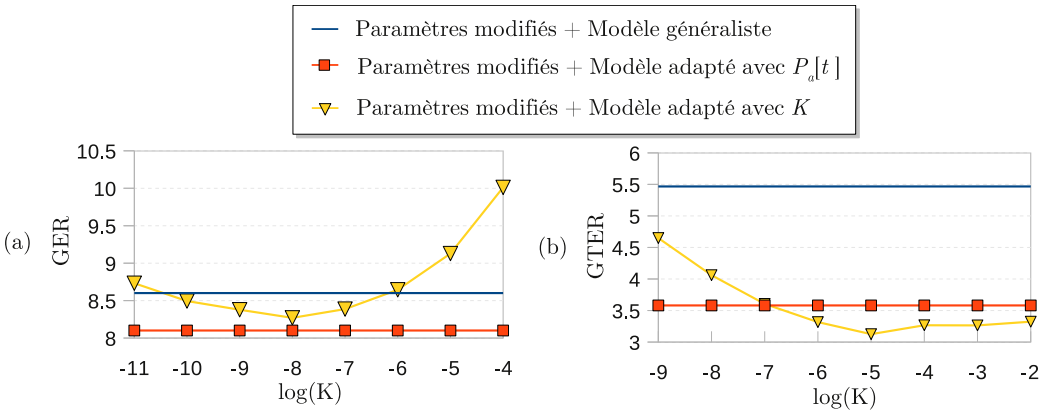


FIGURE 6.4 – GER (a) et GTER(b) obtenus avec le modèle généraliste, les modèles adaptés à partir de $P_a[t]$ ou à partir de différentes valeurs de K.

nouveaux réglages. Nous pensons que ceci est dû à la probabilité linguistique trop faible que peut parfois attribuer notre technique d'adaptation à certains termes simples.

6.5.2 Surpondérer les termes

Comme évoqué en 6.3, la probabilité $P_a[t]$ d'un terme t estimée à partir d'un corpus thématique peut être sous-estimée par rapport à son importance au sein du thème en raison du problème d'éparsité des données. Ceci peut conduire à une adaptation trop faible de certains n-grammes et inhiber l'apparition d'hypothèses de transcription contenant ce terme dans les graphes de mots. Afin d'étudier l'importance des probabilités $P_a[t]$ pour l'adaptation du modèle de langue utilisé au moment de la création des graphes de mots, nous proposons alors d'utiliser un poids K arbitrairement fixé et identique pour tous les termes retenus pour l'adaptation du modèle de langue. Dans ce cas, le facteur de mise à l'échelle $\alpha(w)$ d'un n-gramme hw de la formule (6.20) se réécrit :

$$\alpha(w) = \begin{cases} \dfrac{K}{P_B[w]} & \text{si } w \text{ est un terme,} \\ 1 & \text{sinon .} \end{cases} \tag{6.27}$$

Les figures 6.4.a et 6.4.b montrent respectivement les résultats sur le GER et sur le GTER obtenus pour différentes valeurs de K. Pour des raisons de lisibilité des courbes, nous ne présentons ces résultats que pour nos réglages modifiés du système[37]. Dans l'ensemble, ces résultats mettent en lumière un paradoxe somme toute logique mais intéressant à souligner : des valeurs élevées de K tendent à dégrader le potentiel des graphes du point de vue du GER alors que celles-ci permettent d'obtenir des performances meilleures en ce qui concerne la seule reconnaissance des termes. On note ainsi que la valeur optimale de K est de 10^{-8} en terme de GER mais de 10^{-5} en terme de GTER.

La table 6.13 présente les taux d'erreur sur les mots (WER) et sur les termes (TER) obtenus à partir des graphes de mots générés par ces différentes configurations et dont le décodage s'est poursuivi classiquement par l'utilisation d'un modèle de langue adapté à partir

[37] Les tendances observées pour les réglages d'origine sont néanmoins très similaires à celles présentées, voire même encore plus accentuées.

	Modèle utilisé pour la création des graphes de mots	WER	TER
Réglages d'origine	Généraliste	21,8	14,9
	Adapté à partir de $P_a[t]$	**21,0** $(-0,8)$	12,0 $(-2,9)$
	Adapté à partir de $K = 10^{-8}$	21,2 $(-0,6)$	12,2 $(-2,7)$
	Adapté à partir de $K = 10^{-5}$	21,3 $(-0,5)$	**11,5** $(-3,4)$
Réglages modifiés	Généraliste	20,9	13,7
	Adapté à partir de $P_a[t]$	**20,5** $(-0,4)$	11,3 $(-2,4)$
	Adapté à partir de $K = 10^{-8}$	20,8 $(-0,1)$	11,2 $(-2,5)$
	Adapté à partir de $K = 10^{-5}$	21,0 $(+0,1)$	**11,1** $(-2,6)$

TABLE 6.13 – WER et TER obtenus pour différents modèles de langue utilisés lors de la création des graphes, puis avec un modèle de langue adapté avec P_a pour le reste du décodage. Entre parenthèses, le type de poids utilisé pour le modèle de langue impliqué dans l'étape de création des graphes de mots.

des $P_a[t]$ estimés sur un corpus d'adaptation. Ces résultats sont donnés pour les deux valeurs de K respectivement optimales en terme de GER (10^{-8}) et en terme de GTER (10^{-5}). Avec peu de surprise, il ressort globalement que l'utilisation de poids fixes n'apporte rien voire dégrade le WER, y compris par rapport aux taux rapportés initialement avec le modèle de langue généraliste. À l'inverse, il apparaît que les meilleurs taux de reconnaissance des termes sont obtenus à partir des graphes générés par les modèles de langue adaptés grâce à un poids K élevé. Notons que, cette fois, ce gain n'est pas lié à une plus grande richesse lexicale des graphes de mots puisque nous avons constaté que plus la valeur de K est élevée, plus les graphes de mots sont petits. Notamment, les graphes de mots générés avec $K = 10^{-5}$ contiennent en moyenne 20 % de mots distincts en moins par rapport à ceux obtenus avec le modèle généraliste ou le modèle adapté d'après $P_a[t]$. Paradoxalement, 15 % des mots contenus dans ces graphes de mots générés avec $K = 10^{-5}$ sont des mots qui n'avaient jamais été observés précédemment à partir des autres configurations. Cette dernière propriété permet de faire apparaître dans les transcriptions des termes qui restaient jusqu'à présent mal transcrits. Par exemple, la table 6.14 montre un exemple de groupe de souffle extrait d'un segment traitant de phénomènes d'exode en Afrique. Dans le document audio, le locuteur de ce groupe de souffle est un journaliste à l'accent africain, accent que modélise mal le modèle acoustique du système. Le mot libéria y était alors mal transcrit car celui-ci était élagué lors de la création du graphe de mots. En utilisant un poids K élevé, le terme survit à l'élagage et aboutit à la bonne transcription. Sur un plan applicatif, cette meilleure reconnaissance des termes peut être intéressante dans le cadre d'une utilisation des transcriptions automatiques à des fins d'indexation de documents audio puisque, comme nous l'avons vu avec le critère *tf-idf*, ces techniques s'intéressent peu aux mots grammaticaux mais accordent une importance particulière aux mots discriminants.

Il reste toutefois regrettable de constater que les taux d'erreur sur les mots ne suivent pas la même tendance que ceux sur les termes. Nous pensons que ce comportement est principalement dû à un effet de bord lié à l'introduction dans l'adaptation MDI de la contrainte de conservation de la masse de probabilité des n-grammes observés (formules (6.11) et (6.12)). Alors que cette contrainte est introduite afin de réduire le temps de calcul des facteurs de normalisation des probabilités conditionnelles adaptées, elle inhibe toute modification

Référence	accès à la frontière du libéria
Modèle généraliste	À SERT la ANTIENNE du DÉLIRE
Modèle adapté avec $P_a[t]$	MERCI à la frontière NOUVELLE
Modèle adapté avec $K = 10^{-8}$	MERCI à la frontière LIBÉRIENNE
Modèle adapté avec $K = 10^{-5}$	MERCI à la frontière libéria

TABLE 6.14 – Exemple d'alignement des graphes d'un groupe de souffle de référence pour différents modèles de langue utilisés lors de la création des graphes.

globale de la distribution d'un modèle de langue et n'en induit que des modifications locales, historique par historique. Lorsque des facteurs de mise à l'échelle $\alpha(t)$ trop forts sont utilisés, par exemple lorsque les valeurs de K ou $P_a[t]$ sont élevées, ces adaptations locales tendent à rendre moins probables les historiques des n-grammes adaptés par rapport à ceux pour lesquels aucune probabilité n-gramme n'a été modifiée. Pour illuster ce problème, considérons les n-grammes h_1t, h_1w et h_2w, où h_1 et h_2 sont deux historiques différents, t est un terme et w un mot ordinaire. Dans ce cas, l'adaptation MDI fondée sur l'hypothèse de conservation de la masse de tous les historiques conduit la probabilité $P[t|h_1]$ à être très largement augmentée et à diminuer d'autant la probabilité $P[w|h_1]$ de manière à laisser inchangée la masse de probabilité de l'historique h_1. De son côté, $P[w|h_2]$ n'est pas réestimée car aucun n-gramme h_2t n'est recensé par le modèle de langue. Ainsi, lorsqu'un groupe de souffle contenant w sera décodé à partir du modèle adapté, l'historique h_1 sera préféré à h_2 quand bien même la séquence de mots qui lui correspond ne se justifie pas linguistiquement. Cette particularité implique qu'il est difficile d'obtenir des transcriptions à la fois correctes sur les portions de signal contenant des mots et expressions liés au thème et sur les portions relevant du langage général. Une solution à ce problème pourrait être de combiner les résultats fournis par le modèle généraliste avec ceux fournis par le modèle adapté. Nous montrons toutefois que cette solution n'est pas aussi simple qu'il n'y paraît.

6.5.3 Fusionner les graphes de mots

Comme le montrent les expériences et observations précédentes, il est difficile de concilier au niveau d'une adaptation d'un modèle de langue les informations que l'on possède sur le thème abordé avec celles plus généralistes d'un modèle initial. Il semble alors intéressant de s'orienter vers une intégration *a posteriori* de celles-ci, notamment *via* la fusion de graphes de mots. Pour cela, nous définissons la fusion de deux graphes de mots comme le graphe déterminisé représentant l'union de l'ensemble des phrases codées par chaque graphe. Nous appliquons cette méthode de fusion entre les graphes générés avec le modèle de langue généraliste et ceux obtenus grâce aux modèles adaptés à partir de $K = 10^{-5}$. L'objectif de cette fusion est de produire des graphes de mots comportant les meilleures hypothèses possibles, que le groupe de souffle à décoder comporte des termes ou qu'il ne corresponde qu'à du langage général.

Le tableau 6.15 présente les résultats GER obtenus à partir de cette stratégie de fusion sur notre ensemble de développement et compare ces taux à ceux obtenus sans fusion à partir des différentes configurations utilisées. Il apparaît clairement que la fusion des graphes de mots produit toujours des gains en terme de GER quelles que soient les configurations mises en jeu. Ce résultat s'explique assez logiquement par le fait que la fusion de deux

Avant fusion

	Réglages d'origine	Réglages modifiés
Modèle généraliste	8,9	8,5
Modèle adapté ($K = 10^{-5}$)	9,3	9,1

Après fusion

2^{nd} décodage \ 1^{er} décodage		Réglages d'origine + Modèle généraliste	Réglages modifiés + Modèle généraliste
Réglages d'origine	+ Modèle généraliste	8,6	7,8
	+ Modèle adapté ($K = 10^{-5}$)	8,3	7,7
Réglages modifiés	+ Modèle généraliste	7,8	8,5
	+ Modèle adapté ($K = 10^{-5}$)	**7,6**	8,1

TABLE 6.15 – GER mesurés après fusion des graphes de mots issus d'un premier décodage avec le modèle généraliste et d'un second décodage avec différents modèles de langue.

graphes de mots résulte en un graphe de mots plus gros, contenant donc plus d'hypothèses de transcription. Plus particulièrement, on peut souligner que les graphes fusionnés sont de meilleure qualité lorsque ceux-ci sont issus d'une fusion avec les graphes de mots « adaptés ». Bien que ceci confirme ce que nous attendions, ce phénomène reste d'autant plus surprenant que, avant fusion, ces derniers graphes avaient un GER nettement moins bons que ceux générés par le modèle généraliste. Toutefois, il semblerait que ces gains soient marginaux par rapport à ceux issus globalement de la combinaison des graphes générés à partir de deux configurations différentes de l'algorithme de décodage (cellule grisée). Nous en concluons qu'il existe une certaine complémentarité entre les réglages d'origine, attribuant une importance relativement élevée à l'acoustique, et ceux modifiés qui accordent plus de poids au modèle de langue. La fusion des graphes de mots permet bien de tirer parti de cette complémentarité alors qu'elle ne le permet pas vraiment en ce qui concerne celle entre nos modèles de langue généraliste et thématiquement adapté.

Malgré le potentiel du mécanisme de fusion des graphes de mots que semblent indiquer les différents gains sur le GER, des expériences ultérieures ont montré que ces gains sont lissés en terme de WER à la sortie du système et que nous ne dépassons pas les gains que nous obtenions déjà sans fusion. Ces résultats quelque peu décevants nous apparaissent cependant cohérents avec le fait que notre méthode de fusion n'aboutit qu'à considérer l'union des hypothèses des graphes fusionnés et n'introduit donc aucune nouvelle hypothèse. Ainsi, quel que soit le modèle de langue utilisé pour décoder les graphes de mots fusionnés, celui-ci privilégiera toujours les mêmes hypothèses qu'il privilégiait déjà sans fusion. Deux aspects seraient à envisager pour une meilleure fusion des graphes de mots. D'une part, il faudrait étudier une méthode de combinaison des graphes plus élaborée qui permettrait notamment d'introduire de nouvelles hypothèses. Il pourrait, par exemple, s'agir d'une stratégie basée sur la combinaison de réseaux de confusion (Li *et al.*, 2002). D'autre part, il faudrait être en mesure d'évaluer les graphes de mots fusionnés d'une manière originale, c'est-à-dire indépendante de celles utilisées pour générer les différents graphes de mots à l'origine de la fusion. Ce second point est un problème beaucoup plus difficile pour lequel nous n'avons pas de réponse simple.

6.6 Bilan du chapitre

Au sein de la problématique d'adaptation thématique d'un modèle de langue généraliste, ce chapitre s'est intéressé tant à la manière de réestimer les probabilités n-grammes du modèle qu'aux moyens de tirer le meilleur parti de ces nouvelles probabilités au sein du processus de transcription, tout en restant dans un cadre complètement non supervisé. Cette étude met en avant différentes contributions.

D'une part, nous avons proposé une nouvelle méthode d'adaptation qui spécialise l'adaptation par minimum d'information discriminante en s'appuyant sur des terminologies extraites automatiquement à partir de corpora thématiques tirés du Web. Plus particulièrement, nous avons montré l'intérêt d'utiliser des termes simples pour adapter un modèle de langue. Sur les segments de notre ensemble de test, notre méthode permet d'aboutir, en moyenne et de manière significative, à des diminutions absolues de $0,7\,\%$ sur le WER et de $1,1\,\%$ sur LER pour une diminution relative de $20\,\%$ de la perplexité. Ces résultats sont d'autant plus intéressants qu'ils mettent en lumière le fait que l'adaptation thématique d'un modèle de langue repose sur l'adaptation d'une faible proportion de n-grammes. Par ailleurs, nous avons étudié la possibilité d'affiner cette adaptation par la prise en compte de termes complexes. Nos résultats montrent que ces termes sont difficilement intégrables au sein d'une adaptation d'un modèle n-gramme. Ce constat s'explique notamment par la difficulté à modéliser toutes les variantes linguistiques d'un terme complexe et par l'absence trop fréquente dans le modèle de langue généraliste des n-grammes leur correspondant.

D'autre part, nous avons montré que l'intégration des modèles de langue adaptés doit se faire au plus tôt au sein du processus de transcription. Bien que ceci implique un temps de calcul non négligeable, cette intégration permet aux modèles de langue adaptés de faire apparaître en sortie du système des termes qui étaient jusqu'alors écartés dès la création des graphes de mots par le modèle généraliste. Nous avons par ailleurs montré que cette intégration précoce est parfois insuffisante pour corriger certaines erreurs de transcriptions sur des termes du thème. Nous avons alors diagnostiqué que ce phénomène est dû à des mauvais réglages de l'algorithme de création des graphes de mots ainsi qu'à une réestimation parfois trop faible de certaines probabilités n-grammes. Il apparaît finalement à travers ces différentes études qu'il est difficile d'obtenir un compromis satisfaisant entre la reconnaissance des termes et une bonne reconnaissance globale des autres mots du langage.

Après cette étude détaillée du problème d'adaptation thématique d'un modèle de langue n-gramme, nous pensons que deux tendances majeures émergent parmi les solutions envisageables pour parfaire l'intégration de connaissances linguistiques liées à un thème. Nous distinguons en effet les solutions qui restent dans le cadre de la modélisation n-gramme classique de celles qui s'en écartent.

Si l'on cherche à conserver uniquement un modèle n-gramme, deux limites principales doivent, à notre sens, être dépassées. D'une part, il est nécessaire de pouvoir ajouter de nouveaux n-grammes dans un modèle de langue existant afin de pouvoir donner une existence réelle à des phénomènes observés ou dérivés de l'observation de corpora thématiques. D'autre part, l'adaptation des probabilités d'un modèle doit se faire au niveau des probabilités jointes et non au niveau des probabilités conditionnelles de chaque historique. Dans l'adaptation MDI, cette simplification est indispensable en terme de temps de calcul si l'on vise à réestimer les probabilités de tous les n-grammes, comme c'est le cas dans le cadre générique de cette adaptation. Mais, dans un cas comme le nôtre où seul un sous-ensemble de ces n-grammes est considéré, l'emploi d'une solution intermédiaire devrait être étudiée.

Malgré tout, l'utilisation de n-grammes présente des limites évidentes telles que leur faible portée et leur absence totale de sémantique. Aussi, l'utilisation d'une autre modélisation semble pertinente pour le problème d'adaptation thématique. Une première frange de solutions pourrait consister à utiliser, en complément d'un modèle de langue n-gramme généraliste, des modèles chargés de compenser le score de certaines hypothèses de transcription sous-estimées au regard du thème considéré. Ces modèles additionnels, indépendants du modèle de langue, pourraient par exemple contraindre le parcours de l'espace de recherche ou le décodage de graphes de mots (Lecouteux *et al.*, 2008). Ces solutions ont l'intérêt de ne pas nécessiter de jeter à la poubelle les différents modèles de langue n-grammes déjà présents dans de nombreux systèmes actuels. Par ailleurs, elles peuvent se justifier d'autant plus que le processus de transcription est d'une complexité telle qu'il est parfois plus difficile d'obtenir des changements par des modifications théoriques de la modélisation linguistique qu'en accédant directement aux rouages du système. L'emploi d'un modèle de langue plus sophistiqué est néanmoins une piste intéressante. Comme nous l'avons vu au chapitre 2, de nombreuses méthodes pour intégrer différents types de connaissances de la langue ont été proposées dans la littérature. Bien que celles-ci ne montrent cependant pas de gains dans l'absolu par rapport aux modèles n-grammes, nous pensons que ces autres modélisations peuvent servir la tâche d'adaptation thématique. Parmi celles-ci, les modèles factoriels, qui modélisent les mots comme des vecteurs de caractéristiques, nous semblent être la modélisation la plus prometteuse. En attachant aux mots du vocabulaire des informations morphologiques et sémantiques, la généralisation des observations faites à partir d'un corpus serait en effet facilitée. Toutefois, cette modélisation repose encore une fois sur une hypothèse markovienne qui réduit finalement la portée des connaissances introduites. Ainsi une dernière technique envisageable pour aller plus loin dans la tâche d'adaptation thématique serait de s'appuyer sur un modèle à amorces. Pour rappel, ce type de modèle prend en compte des relations entre mots distants dans le discours grâce à un cache et à des relations de déclenchement entre mots. Contrairement à l'approche standard qui s'appuie sur les mêmes relations quel que soit le thème rencontré, il serait intéressant chercher à ajouter des relations dynamiquement, en fonction du thème rencontré. Il ne faut toutefois pas perdre de vue qu'un inconvénient majeur de cette dernière option est que la prise en compte de relations distantes se concilie mal avec la mise en place d'un algorithme de décodage efficace. Ce dernier serait en effet particulièrement ralenti par le nombre considérable de caches à considérer pour chaque hypothèse de transcription possible. Ainsi, au delà des problèmes théoriques, de véritables aspects algorithmiques seraient donc également à traiter.

Chapitre 7

Adaptation du vocabulaire

Ce chapitre aborde le dernier volet de notre chaîne d'adaptation thématique d'un système de reconnaissance automatique de la parole à travers la tâche d'adaptation de son vocabulaire. Dans l'absolu, cette tâche vise à apprendre au système de nouveaux mots susceptibles d'être prononcés dans un document traitant d'un thème particulier. Cette adaptation s'avère capitale pour deux raisons. D'une part, certains mots hors vocabulaire sont des mots importants pour le thème, par exemple, des noms de personnes ou des termes techniques, et il est donc primordial de reconnaître ces mots si l'on cherche à exploiter le contenu sémantique d'un document audio. D'autre part, les erreurs de transcription produites par le système pour chaque mot hors vocabulaire d'un document à transcrire tendent à se répercuter sur d'autres mots environnants, ce qui engendre typiquement plusieurs erreurs de transcription pour chaque occurrence d'un mot hors vocabulaire. L'adaptation d'un vocabulaire se décompose en deux étapes. Il faut en effet déterminer quels sont les mots à ajouter au vocabulaire, puis intégrer ces nouveaux mots dans le système, c'est-à-dire dans son lexique phonétisé et son modèle de langue.

Nous nous sommes intéressé à ces différents problèmes durant la fin de notre thèse. Ce chapitre, plus court, présente donc un caractère plus exploratoire et n'expose une solution que partiellement expérimentée. Par ailleurs, nous limitons notre problème à l'étude des mots hors vocabulaire prononcés dans un segment donné, le but étant de retrouver ces mots dans un corpus thématique extrait d'Internet et de les intégrer au mieux dans le modèle de langue généraliste de notre système afin d'améliorer la transcription automatique du seul segment. Plus précisément, conformément à la problématique générale de cette thèse, nous nous sommes focalisé, parmi les mots hors vocabulaire d'un segment, sur la reconnaissance de ceux liés au thème abordé. En ce sens, nous n'apportons que peu d'importance aux noms propres car ceux-ci réfèrent généralement une entité unique (une personne, un lieu, *etc.*) et sont donc difficilement associables à un concept sémantique dont on pourrait dire s'il est spécifique ou non au thème d'un segment. Par ailleurs, ces mots ont la caractéristique d'avoir des propriétés linguistiques particulières comme le fait d'être fréquemment composés de plusieurs mots et d'être souvent accompagnés de marques telles qu'un titre de civilité. Pour cette raison, l'emploi d'une ou plusieurs classes leur étant spécialement dédiées au sein d'un vocabulaire ouvert se présente déjà comme une solution probante pour traiter les noms propres hors vocabulaire. Certains noms propres font néanmoins figure d'exception en se caractérisant par un emploi substantif, c'est-à-dire accompagné d'un déterminant. Cette caractéristique vaut, par exemple, pour les sigles (« la CGT » pour « la Confération Générale

		Ensemble de développement	Ensemble de test
	Taux de mots hors vocabulaire	$0,9\,\%$	$1,1\,\%$
Tous les mots hors vocabulaire	Nombre d'occurrences	310 $(3,4)$	340 $(4,2)$
	Nombre de formes	252 $(2,8)$	257 $(3,2)$
Mots communs thématiques	Nombre d'occurrences	91 $(1,0)$	78 $(1,0)$
	Nombre de formes	83 $(0,9)$	72 $(0,9)$

TABLE 7.1 – Statistiques sur les mots hors vocabulaire dans nos ensembles de développement et de test. Entre parenthèse, les valeurs moyennes par segment.

des Travailleurs »), les marques utilisées pour désigner la propriété essentielle d'une classe d'objets (« un M16 » pour « un fusil de type M16 ») ou d'autres noms propres pour lesquels des raccourcis linguistiques sont devenus d'usage courant (« l'A1 » pour « l'autoroute A1 »). Pour résumer, dans ce chapitre, nous nous intéressons donc aux mots hors vocabulaire qui sont présents dans la transcription de référence d'un segment donné et qui sont importants pour le thème de ce segment, à l'exclusion des noms propres n'ayant pas valeur de substantif. Dans la suite, nous simplifions nos propos par l'usage du terme « mots communs thématiques » pour référer ces mots.

Ces précisions données, nous commençons par étudier la possibilité de repérer les mots communs thématiques à ajouter au système à partir de nos corpora thématiques. Nous montrons notamment que ces corpora ont un potentiel intéressant mais qu'ils comportent beaucoup de bruit. Nous suggérons alors une méthode de filtrage visant à diminuer ce bruit. Nous nous intéressons ensuite à l'intégration de nouveaux mots dans un modèle de langue et proposons une méthode originale qui permet de déterminer automatiquement les n-grammes au sein desquels un nouveau mot est susceptible d'apparaître. En ajoutant ces n-grammes dans un modèle de langue pré-existant, nous montrons que cette technique produit d'aussi bons résultats que les méthodes de l'état de l'art fondées sur un vocabulaire ouvert. Pour valider notre technique, nous ne modifions, au cours de nos expériences, que le vocabulaire du système et n'adaptons pas les probabilités n-grammes de notre modèle de langue généraliste initial.

7.1 Recherche de mots à ajouter

Typiquement, les mots à ajouter au vocabulaire d'un système sont sélectionnés parmi une liste de mots dont on sait qu'ils sont absents du vocabulaire. Comme nous l'avons vu au chapitre 3, cette liste peut être issue de diverses ressources telles que des lexiques fournis par des experts ou des corpora textuels. La difficulté de cette tâche est alors de savoir diagnostiquer les mots les plus susceptibles d'apparaître pour un document donné à transcrire d'après son signal sonore, son genre, son thème ou encore sa période. La philosophie de cette thèse voudrait que nous recherchions alors l'ensemble des mots hors vocabulaire liés au thème d'un segment donné. Le problème est cependant que cette dernière recherche se solderait probablement par un très grand nombre de mots à ajouter, notamment eu égard au très grand nombre de flexions que peuvent supporter certains mots. Or, il n'est clairement pas envisageable d'ajouter trop de nouveaux mots au vocabulaire initial du système car ceux-ci augmenteraient considérablement l'ambiguïté du modèle de langue et du

Thème : crash, avion	
aspirées	✓
iliouchine[38]	✓
kasonga	
kibambi	
mi-parcours	✓
ndjili	
nonguillonof	
shintwa	
transall	

Thème : procès, Irak	
al-majid	
al-tikriti	
brié	
gazé	✓
janton	
statuaient	✓

Thème : rhume, grippe	
clouent	✓
contaminent	✓
éternuant	✓
guériront	✓
influenza	✓
grippés	✓
orthomyxovirus	✓
nasales	✓
mouchant	✓
rhinovirus	✓

TABLE 7.2 – Liste mots hors vocabulaire pour trois segments traitant de thèmes variés. Les noms propres sont en gras et les mots communs thématiques sont marqués d'une coche.

modèle acoustique. Dans notre cas précis, nous cherchons donc uniquement à déterminer l'ensemble des mots communs thématiques effectivement présents dans la transcription de référence d'un segment donné à partir de son corpus thématique de pages Web. Cette section montre tout d'abord que, en dépit du potentiel de ce corpus, la nature bruitée de ses textes rend difficile cette tâche de sélection. Nous exposons alors notre suggestion quant à la manière résoudre ce problème de bruit.

7.1.1 Statistiques sur les mots hors vocabulaire

Le tableau 7.1 présente les taux de mots hors vocabulaire au sein de nos segments thématiques, c'est-à-dire la proportion des occurrences de mots hors vocabulaire dans leur transcription de référence. Les nombres concrets d'occurrences de mots et de formes distinctes que représentent ces taux sont par ailleurs fournis pour nos ensembles de développement et de test. Globalement, il ressort qu'environ 1 mot prononcé sur 100 est un mot inconnu du système. Plus en détail, ce taux signifie que chacun de nos segments thématiques contient en moyenne 3 ou 4 occurrences de mots hors vocabulaire et que ces quelques occurrences correspondent à environ 3 mots différents. Si l'on s'intéresse aux seuls mots communs thématiques, on s'aperçoit que ces derniers ne représentent qu'environ 1 mot hors vocabulaire sur 3. Cette proportion est illustrée par les mots hors vocabulaire extraits de trois segments thématiques différents, présentés dans la table 7.2. Les deux premiers cas sont des exemples typiques. On y remarque qu'une majeure partie des mots hors vocabulaire sont des noms propres, le reste se constituant principalement de flexions de verbes (par exemple statuaient) ou des variantes compositionnelles de mots (par exemple mi-parcours ou orthomyxovirus) absentes du lexique. Toutefois, comme en témoigne notre troisième exemple, il existe une grande variabilité entre segments. Certains segments thématiques ne visent en effet pas à relater un fait d'actualité mais présentent un travail de fond sur un domaine donné, par exemple littéraire ou scientifique. Ces segments comportent alors une abondance de termes inconnus du système pour seulement très peu de noms propres. Par ailleurs, soulignons que le nombre de mots hors vocabulaire présents d'un segment à l'autre est extrêmement variable. Ceci s'explique par le fait que, dès lors que le thème d'un seg-

[38] L'*Iliouchine* est un avion russe utilisé comme un nom commun dans le segment en question.

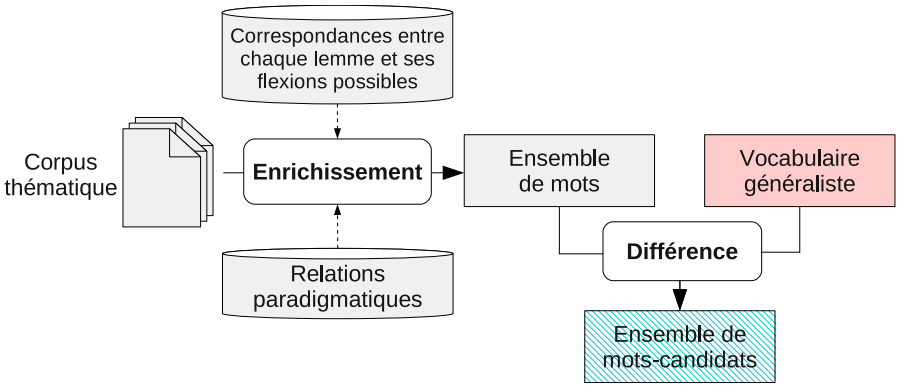


FIGURE 7.1 – Constitution d'une liste de termes-candidats pour l'adaptation thématique du vocabulaire.

ment n'a que rarement, voire jamais, été abordé dans les textes du corpus d'apprentissage du système, la couverture du vocabulaire initial est pauvre pour ce thème et le nombre de mots hors vocabulaire pour le segment est élevé. Au contraire, s'il s'agit d'un thème plus fréquent, cette couverture est nettement meilleure. Le problème premier dans l'adaptation d'un vocabulaire est donc de combler cette faiblesse en recherchant de nouveaux mots liés au thème.

7.1.2 Potentiel des corpora thématiques

Pour trouver des mots hors vocabulaire liés à un thème, une idée assez commune consiste à utiliser le corpus thématique attaché à chaque segment. Des expériences préliminaires ont en effet montré qu'un corpus thématique, tel que nous le construisons au chapitre 5, intègre en moyenne 35% des mots hors vocabulaire d'un segment donné. Ce pourcentage signifie que, parmi les 3 mots hors vocabulaire moyens d'un segment, seulement l'un d'entre eux est présent dans le corpus thématique. Si l'on ne s'intéresse qu'aux mots communs thématiques, ce pourcentage s'améliore et monte à 52 %. Malheureusement, il apparaît que 39 % des formes présentes dans chaque corpus thématique sont des mots hors vocabulaire, soit environ 11 400 mots par corpus. Concrètement, ceci signifie donc qu'il s'agit en moyenne de retrouver un seul mot parmi un ensemble de 11 400 mots-candidats ! Ce nombre élevé s'explique en majeure partie par le fait que les pages Web à l'origine de nos corpora thématiques contiennent beaucoup de noms propres, de mots étrangers (notamment anglais) et de fautes d'orthographe. Outre la dégradation des modélisations acoustiques et linguistiques qu'impliquerait l'inclusion de ces nombreux mots, ceux-ci conduiraient également le système à produire des transcriptions contenant de nombreuses fautes d'orthographe. L'utilisation directe des corpora thématiques n'est donc pas une solution réaliste. Nous suggérons alors une méthode permettant, d'une part, d'augmenter la proportion de mots communs thématiques hors vocabulaire à retrouver à partir du corpus thématique de chaque segment et, d'autre part, de filtrer efficacement l'ensemble des mots-candidats, limitant ainsi le nombre de mots à ajouter au vocabulaire du système.

7.1.3 Processus de recherche de mots communs thématiques

Pour dépasser les limites liées à l'utilisation directe des corpora thématiques, nous suggérons une stratégie en deux temps. Une première étape cherche à généraliser le contenu

	Mots du corpus uniquement	Mots du corpus + toutes leurs flexions	Mots du corpus + flexions + relations paradigmatiques
Pourcentage de mots communs thématiques retrouvés pour chaque segment	52 %	64 %	70 %
Nombre de mots-candidats par corpus thématique	11 400	35 600	40 600

TABLE 7.3 – Statistiques sur les ensembles de mots-candidats dérivés des mots des corpora thématiques grâce à différentes techniques d'enrichissement.

des corpora thématiques de manière à augmenter les chances d'inclure des mots pertinents pour le thème considéré alors qu'une seconde étape vise à filtrer thématiquement ce contenu généralisé de manière à ne pas ajouter trop de nouveaux mots dans le vocabulaire d'origine du système.

Lorsqu'ils ne sont pas des noms propres, les mots hors vocabulaire peuvent être des flexions rares de mots eux-mêmes rares. Dans des expériences préliminaires, nous avons cherché à savoir s'il était possible de retrouver ces flexions à partir des lemmes des mots observés dans les corpora thématiques ainsi que d'informations flexionnelles et paradigmatiques. Comme l'illustre la figure 7.1, nous avons complété les mots rencontrés dans un corpus par toutes les flexions possibles de leur lemme et des lemmes des mots avec lesquels ils partagent une relation paradigmatique issue d'un apprentissage automatique[39]. La table 7.3 présentent différentes statistiques concernant la liste des mots-candidats générés à partir de ces différents enrichissements pour chaque corpus thématique. En utilisant les informations flexionnelles et paradigmatiques des mots des corpora thématiques, ces résultats montrent que 70 % des mots communs thématiques hors vocabulaire présents dans les segments thématiques peuvent en moyenne être retrouvés pour chaque segment. En contrepartie cependant, le nombre de mots-candidats à considérer pour chaque adaptation du vocabulaire est multiplié par 4.

Si l'on cherche à maximiser nos chances de retrouver les mots hors vocabulaire prononcés dans un segment thématique, il faut donc mettre l'accent sur une stratégie de filtrage des très vastes listes de mots-candidats générées à partir des corpora thématiques. Bien que cette question nous intéresse, nous ne l'avons pas abordée plus avant. Nous pensons toutefois qu'il est indispensable d'effectuer ce filtrage par la combinaison de critères complémentaires. D'une part, il serait intéressant de développer une mesure d'importance thématique d'un mot-candidat par rapport à un segment considéré. Cette mesure pourrait par exemple s'effectuer à partir des mêmes outils que nous avons utilisés pour la création de terminologies, à la condition près qu'il faudrait probablement accorder plus de poids

[39]Cet apprentissage a été affiné en comparaison de celui qui avait été mis en œuvre pour nos expériences liées à l'adaptation du modèle de langue (*cf.* section 6.3). Plus précisément, chaque lemme d'un vaste corpus textuel (en l'occurrence, les archives du journal *Le Monde* sur 20 ans) est associé à un vecteur contenant le nombre de fois où ce lemme apparaît en compagnie des autres lemmes du texte dans une fenêtre de 10 mots à gauche et à droite. La distance *cosinus* entre les vecteurs de contexte de deux lemmes ℓ_1 et ℓ_2 permet de ne conserver, pour chaque lemme ℓ_1, que les 10 lemmes ℓ_2 partageant les liens les plus forts. Par ailleurs, la valeur cosinus donne une indication du degré d'affinité entre mots parmi les liens conservés.

aux mots rares car ce sont des mots de cette nature que nous recherchons. D'autre part, l'utilisation d'un filtrage des mots-candidats à partir d'un critère acoustique semble incontournable si l'on ne veut pas avoir à ajouter un trop grand nombre de mots au vocabulaire initial du système à chaque adaptation. En alignant par exemple des transcriptions phonétiques de chaque mot-candidat avec celle du segment thématique, il serait ainsi possible de ne conserver que les mots qui, acoustiquement parlant, semblent avoir été prononcés dans le segment considéré.

Une fois déterminés les mots à ajouter au vocabulaire du système, il reste encore à ajouter ces mots dans le lexique phonétisé et à faire en sorte qu'ils soient pris en compte par le modèle de langue du système.

7.2 Intégration de nouveaux mots

L'intégration de nouveaux mots dans un système de transcription pré-existant comprend deux étapes. Dans un premier temps, il faut ajouter ces mots dans le lexique phonétisé du système de manière à leur donner une existence acoustique. Comme nous l'avons vu au chapitre 2, cette étape peut être réalisée via l'utilisation d'outils de phonétisation automatique. Nous ne nous intéressons ainsi pas à cette tâche. Dans un second temps, il faut que les mots à ajouter soient pris en compte par le modèle de langue de manière à pouvoir calculer la probabilité des hypothèses de transcription contenant ces nouveaux mots. Une approche intéressante déjà développée pour résoudre ce problème consiste à créer un modèle de langue à vocabulaire ouvert (Allauzen et Gauvain, 2005b). En incluant des classes ouvertes représentant, par exemple, des catégories morphosyntaxiques parmi le reste des mots habituels du vocabulaire, cette technique consiste à ajouter chaque nouveau mot comme une variante de prononciation de la classe qui lui correspond. Nous adressons cependant à cette méthode le reproche d'associer à chaque mot d'une classe le même rôle sémantique.

Dans cette section, nous cherchons à ajouter directement des n-grammes contenant les nouveaux mots du vocabulaire dans un modèle de langue au vocabulaire fermé. À cette fin, il faut déterminer dans quels contextes syntaxiques apparaissent les nouveaux mots afin d'en dériver des n-grammes, puis estimer les probabilités conditionnelles de chacun de ces nouveaux n-grammes. Pour cela, nous proposons d'associer chaque nouveau mot à un ensemble de mots qui leur sont liés sémantiquement et qui sont déjà présents dans le vocabulaire initial. De cette manière, les n-grammes liés à un nouveau mot sont dérivés à partir des n-grammes contenant ces mots sémantiquement voisins. Nous présentons tout d'abord notre méthode de recherche de nouveaux n-grammes et traitons ensuite le problème de l'estimation de leur probabilité. Enfin, nous présentons des résultats confrontant notre méthode à d'autres techniques existantes, notamment celle fondée sur un vocabulaire ouvert.

7.2.1 Constitution des *n*-grammes d'un nouveau mot

Au sein d'une langue, différentes informations régissent l'utilisation d'un mot dans une phrase. D'une part, le choix d'un mot est dicté par le sens que celui-ci partage avec les autres mots. Cette information sémantique favorise, par exemple, l'utilisation de certains adjectifs ou verbes pour accompagner un mot donné. Ainsi, certains mots peuvent se substituer les uns aux autres sans que cela n'altère l'intelligibilité d'une phrase, alors que, au

contraire, d'autres substitutions ôtent tout sens à une expression. Par exemple, dans la phrase « Jacques mastique un chewing-gum. », le fait de remplacer le verbe mastiquer par machouiller ou manger produit des énoncés intelligibles, bien que le sens change toutefois légèrement. À l'inverse, si l'on remplace mastiquer par numéroter, l'énoncé perd tout sens. D'autre part, comme le langage répond à des règles grammaticales, l'emploi valide d'un mot dépend de son information morphosyntaxique, c'est-à-dire de sa catégorie grammaticale et de sa flexion.

À partir de ce constat, notre idée pour déterminer les n-grammes au sein desquels peut apparaître un nouveau mot consiste à assimiler ce mot à d'autres partageant une utilisation grammaticale et sémantique commune, puis à rechercher les n-grammes au sein desquels apparaissent ces mots équivalents. Plus formellement, nous associons à chaque nouveau mot w son lemme ℓ_w et sa catégorie morphosyntaxique m_w[40]. En considérant une relation sémantique $\mathcal{R}$ entre lemmes, nous posons alors comme équivalents deux mots dont les lemmes sont en relation d'après $\mathcal{R}$ et qui partagent une même catégorie morphosyntaxique :

$$w \equiv v \quad \Leftrightarrow \quad \ell_w \mathcal{R}\, \ell_v \text{ et } m_w = m_v, \tag{7.1}$$

où $\ell_w \mathcal{R}\, \ell_v$ signifie que ℓ_w est en relation avec ℓ_v. Cette notion d'équivalence entre mots nous permet de définir deux n-grammes comme équivalents s'ils ne diffèrent que par la substitution d'un mot en un autre qui lui est équivalent.

La recherche des n-grammes relatifs à un nouveau mot s'exprime alors comme celle des n-grammes les plus équivalents. Pour cela, étant donné un mot à ajouter w, nous évaluons l'importance d'un n-gramme XwY par un score $S(XwY)$ qui mesure la fréquence d'apparition des n-grammes équivalents à XwY parmi un vaste ensemble E de n-grammes :

$$S(XwY) = \sum_{w \equiv v} r(\ell_w, \ell_v) \times |XvY|_E, \tag{7.2}$$

où $r(\ell_w, \ell_v)$ est le score du lien entre ℓ_w et ℓ_v et $|XvY|_E$ désigne le nombre de fois où le n-gramme XvY apparaît dans E. En pratique, nous considérons E comme la concaténation des n-grammes extraits de notre corpus généraliste avec ceux observés dans un corpus thématique. À l'issue de cette étape, nous calculons, pour chaque ordre n, le score moyen des n-grammes ayant un score S non nul. Finalement, seuls les n-grammes dont le score est supérieur au score moyen de leur ordre sont conservés. Pour mettre en œuvre cette méthode, il est théoriquement nécessaire d'évaluer tous les n-grammes possible. Ceci requierrait un temps de calcul exponentiel en fonction de n, problème d'autant plus gênant que, bien que ce ne soit pas le cas dans nos expériences, le nombre de mots à ajouter pour un segment peut être important. Toutefois, cette complexité peut être largement réduite par un parcours séquentiel opéré une seule fois sur l'ensemble E des n-grammes en s'appuyant sur la relation réciproque $\mathcal{R}^{-1}$.

Dans nos expériences, nous avons instancié la relation $\mathcal{R}$ par des relations paradigmatiques apprises automatiquement. Ces relations indiquent à quel point deux lemmes partagent des contextes d'apparation similaires (Grefenstette, 1994). La table 7.4 donne des exemples de trigrammes que nous déterminons alors automatiquement pour les mots hors vocabulaire jamaïcaines, cancérologues et garent. Nous précisons par ailleurs les lemmes équivalents ayant permis d'obtenir ces résultats. Il apparaît que les n-grammes construits ont du sens et sont grammaticalement corrects. On note toutefois que, dans certains cas,

[40] Pour les verbes conjugués, nous décidons de supprimer l'information concernant le mode et le temps.

JAMAÏCAINES (afro-américain, antillais, métissé, pop, techno)	CANCÉROLOGUES (biologiste, cardiologue, chirurgien, docteur juriste, neurologue, pédiatre, psychiatre)	GARENT (immobiliser, incendier, percuter, stationner)
les musiques jamaïcaines	cancérologues spécialisés dans	garent sur la
ses origines jamaïcaines	cancérologues des hôpitaux	garent sur le
jamaïcaines et africaines	de nombreux cancérologues	garent devant la
musiques jamaïcaines et	de cancérologues américains	garent le long
quartier des jamaïcaines	pour les cancérologues	garent en permanence
des chansons jamaïcaines	plupart des cancérologues	garent en double
ses racines jamaïcaines	les cancérologues estiment	jeunes qui garent
des jamaïcaines non	cancérologues et médecins	garent devant le
les eaux jamaïcaines	selon les cancérologues	qui y garent
jamaïcaines et caraïbe	les cancérologues ont	caravanes garent sur

TABLE 7.4 – Exemple de n-grammes construits automatiquement. Entre parenthèses, le lemme des mots équivalents d'après nos relations paradigmatiques.

des traitements additionnels *a posteriori* seraient bienvenus. Par exemple, les n-grammes relatifs au mot garent sont principalement construits à partir de l'existence d'une relation paradigmatique entre garer, qui peut s'employer sous une forme pronominale, avec des verbes qui, eux, sont non pronominaux. Il en découle des n-grammes incorrects tels que « caravanes garent sur » au lieu de « caravanes se garent sur ».

Après avoir observé l'ensemble de nos résultats, il apparaît que certains de nos mots hors vocabulaire sont des mots composés et que, étant donné la rareté de ces mots dans un corpus, notre technique d'apprentissage de relations paradigmatiques ne permet, la plupart du temps, pas de lier ces mots à d'autres. Ceci implique qu'aucun n-gramme équivalent ne peut être trouvé et que, par exemple, un mot hors vocabulaire comme rhinovirus n'aura que son seul unigramme dans le modèle de langue final. Pour pallier ce problème, il est envisageable de réduire les mots composés hors vocabulaire en une forme plus simple (par exemple, la forme virus). Bien que n'ayant pas à disposition d'outil permettant d'effectuer une telle réduction de manière automatique, nous avons voulu mesurer le potentiel que représenterait la connaissance de ces relations « composé-composant » entre mots – nous parlons dans la suite de relations compositionnelles. Pour cela, nous avons manuellement généré les formes de base de nos mots hors vocabulaire composés, par exemple « rhinovirus → virus ». À titre indicatif, ces mots sont au nombre de 24 dans notre ensemble de développement et de 20 dans notre ensemble de test, soit respectivement 29 % et 28 % de l'ensemble des mots communs thématiques que nous considérons.

Maintenant que nous avons déterminé un ensemble de n-grammes à ajouter dans le modèle de langue, il nous reste à calculer les probabilités conditionnelles de ces n-grammes et à mesurer leur effet en terme de diminution du taux d'erreur.

7.2.2 Intégration dans un modèle de langue et résultats

Étant donné un ensemble de n-grammes basé sur de nouveaux mots, nous devons calculer la probabilité conditionnelle de chacun des n-grammes et intégrer ces probabilités dans le modèle de langue du système. À cette fin, pour chaque nouveau n-gramme hw, nous calculons tout d'abord la probabilité jointe $P[hw]$ comme la moyenne des probabilités jointes des n-grammes équivalents à hw, puis nous ramenons cette moyenne vers une probabilité

conditionnelle. Dans le cas toutefois où aucun n-gramme équivalent n'a pu être trouvé pour un nouveau mot, seule une probabilité unigramme fixée arbitrairement à 10^{-8} est utilisée. Nous n'avons pas cherché à optimiser cette probabilité par défaut étant donné que, dans les segments de notre corpus de développement, le recours à celle-ci n'a lieu que pour 19 % des mots hors vocabulaire lors de l'utilisation de relations paradigmatiques seules et 9 % lorsque les relations compositionnelles sont ajoutées. Ensuite, nous ajoutons les nouveaux n-grammes et leur probabilité conditionnelle pour chaque historique concerné dans le modèle de langue, nous mettons à jour la masse de probabilité des évènements observés de ces historiques, puis nous recalculons leur score de repli (*backoff*).

Après avoir lancé un nouveau processus de transcription à partir des vocabulaires et modèles de langue enrichis, nous mesurons les WER sur les segments thématiques qui contiennent au moins un mot commun thématique hors vocabulaire. Ces segments sont respectivement au nombre de 42 pour l'ensemble de développement et de 37 pour l'ensemble de test. Nous comparons ces taux avec ceux théoriquement atteignables si tous les mots commun thématiques hors vocabulaire ajoutés au vocabulaire étaient bien transcrits par le système généraliste – nous parlerons de taux *oracle* – ainsi que ceux obtenus pour deux autres méthodes que nous avons implémentées. D'une part, nous comparons notre méthode à une technique minimaliste qui consiste à n'ajouter dans le modèle de langue initial que les unigrammes correspondant à chaque mot ajouté au vocabulaire. Cette technique s'appuie alors sur le seul espoir que ces mots soient suffisamment peu ambigus sur le plan phonétique pour qu'ils soient reconnus d'après le seul modèle acoustique. D'autre part, nous comparons notre méthode à celle fondée sur un modèle de langue à vocabulaire ouvert incluant 14 classes grammaticales (verbe, nom commun, adjectif...). Le modèle initial utilisé pour cette méthode a été construit en suivant le même exact processus d'apprentissage que celui ayant servi à construire notre modèle de langue à vocabulaire fermé. Nous avons notamment fait en sorte que les WER par défaut de ces deux modèles initiaux soient sensiblement les mêmes, que ce soit pour notre ensemble de développement ou pour celui de test. Enfin, pour commenter la comparaison des résultats des techniques mises en œuvre, nous nous appuyons sur quelques exemples de transcription contenus dans la table 7.5.

La table 7.6 présente les WER obtenus pour les modélisations à vocabulaire fermé et à vocabulaire ouvert, avant et après enrichissement du vocabulaire. Concernant notre méthode, nous présentons notamment les résultats obtenus par l'utilisation de relations paradigmatiques seules (*par*) et par l'utilisation complémentaire de relations de composition (*par+comp*). Tout d'abord, ces résultats montrent que pour un gain maximal théorique sur le WER de $0,6$, l'utilisation de seuls unigrammes permet déjà d'atteindre quasi la moitié de ce gain. Comme nous le soulignions plus haut, ceci s'explique par le fait que les mots hors vocabulaire sont généralement des mots longs et qu'ils sont donc plutôt bien reconnus par le modèle acoustique (exemple 1). Par ailleurs, il apparaît que l'intégration de n-grammes grâce aux relations paradigmatiques seules produit des gains encore meilleurs. Ces gains sont similaires à ceux obtenus par l'utilisation du modèle de langue à vocabulaire ouvert, gains tout deux statistiquement significatifs[41]. Ceci souligne la qualité des n-grammes qui nous repérons automatiquement. Toutefois, nous notons que l'adjonction de relations de composition n'apporte rien voire conduit à des gains plus faibles. Nous expliquons ceci par le fait que, après analyse des transcriptions, les mots composés sont parmi les plus simples à transcrire étant donné leur longueur et, donc, leur faible ambiguïté acoustique (exemple 1

[41] Pour les tests de Student et de Wilcoxon, les valeurs p valent respectivement $2,7 \times 10^{-5}$ et $2,2 \times 10^{-5}$ pour notre technique, et toutes deux $1,7 \times 10^{-6}$ pour le modèle de langue à vocabulaire ouvert.

à nouveau). Ainsi, ajouter des n-grammes supplémentaires n'apporte rien. Au contraire, il semble même dans certains cas que ces nouveaux n-grammes perturbent les probabilités conditionnelles de certains historiques du modèle de langue, ce qui aboutit à des erreurs qui n'existaient pas auparavant sur des zones où aucun mot hors vocabulaire n'est pourtant présent. Par ailleurs, il semblerait que la réduction des mots composés vers leur composante de tête ne soit pas toujours la bonne solution pour traiter un mot hors vocabulaire de ce genre. Dans l'exemple 2, le mot intercongolais est ramené par notre méthode vers le mot congolais afin de construire de nouveaux n-grammes. Or, comme le montre la référence du groupe de souffle, il apparait que c'est le sens qu'apporte le préfixe inter, ici pour désigner des relations mutuelles, qui permet l'emploi de intercongolais avec le nom dialogue. Il aurait ainsi été préférable de ramener intercongolais vers des mots tels que international ou intercommunautaire. Aussi, d'autres pistes d'emploi des relations de composition devraient être étudiées.

Pour juger de l'efficacité de notre méthode pour reconnaître les mots communs thématiques ajoutés au système, nous mesurons les WER uniquement sur ces derniers. Nous notons cette mesure WER_{HV} pour distinguer ce taux du WER classique, sur l'ensemble des mots. Nous présentons ces scores dans la table 7.7 pour les mêmes configurations que celles utilisées pour le calcul du WER classique. Ces résultats laissent transparaître les mêmes tendances que précédemment entre méthodes fondées uniquement sur des unigrammes ou sur des n-grammes. Toutefois, l'analyse des autres résultats conduit à un bilan légèrement différent de celui que nous pouvions faire d'après le WER classique. En effet, il apparaît que notre méthode aboutit à un WER_{HV} moins bon que celui obtenu par le modèle de langue à vocabulaire ouvert alors que nous obtenons des gains similaires pour le WER classique. Ceci signifie, certes, qu'un mot hors vocabulaire a moins de chance d'être correctement transcrit par notre technique mais que, lorsqu'il l'est, la transcription des mots qui l'entourent est meilleure que celle fournie par un modèle de langue à vocabulaire ouvert. Cette conclusion est illustrée par l'exemple 3 dans lequel notre méthode a permis d'associer le mot djellabas avec d'autres noms de vêtements et de prendre ainsi en compte l'association avec le mot bleues alors que le modèle de langue à vocabulaire ouvert considère comme plus probable une hypothèse dépourvue de sens. Notre technique conduit donc à mieux intégrer de nouveaux mots dans les contextes d'utilisation qui leur correspondent.

Finalement, ces différents résultats soulignent que, globalement, notre méthode fournit des résultats satisfaisants mais que des efforts devraient être faits pour affiner certains points. D'une part, au delà de l'information paradigmatique, nous pensons qu'il serait bon d'utiliser d'autres types de relations entre mots pour mettre en œuvre notre recherche de n-grammes équivalents. L'utilisation conjointe de ces relations pourrait alors se faire par une combinaison des n-grammes produits pour chaque relation. Par exemple, lorsque le nombre de n-grammes retenus pour un mot à ajouter est trop faible à partir de la seule information paradigmatique, il pourrait être intéressant de chercher à compléter ceux-ci en associant le mot à n'importe quel autre mot partageant la même classe grammaticale. D'autre part, la détermination des probabilités des n-grammes ajoutés au modèle de langue est un problème difficile. Une attention particulière devrait notamment être portée à la normalisation des probabilités après enrichissement car, au moment de construire le modèle initial, cette étape est issue de calculs complexes impliquant des techniques de lissage ainsi que le calcul de scores de repli.

Exemple 1

Référence		ce sont des mouvements **panislamistes**
Sans enrichissement		ce sont des mouvements PAGNE islamiste
Vocabulaire fermé	+ unigrammes uniquement	ce sont des mouvements panislamistes
	+ n-grammes (par)	ce sont des mouvements panislamistes
	+ n-grammes ($par + comp$)	ce sont des mouvements panislamistes
Vocabulaire ouvert		ce sont des mouvements panislamistes

Exemple 2

Référence		le comité de suivi du dialogue **intercongolais**
Sans enrichissement		le comité de suivi ÉVIDEMMENT UN DES CONGOLAIS
Vocabulaire fermé	+ unigrammes uniquement	le comité de suivi ÉVIDEMMENT **intercongolais**
	+ n-grammes (par)	le comité de suivi ÉVIDEMMENT **intercongolais**
	+ n-grammes ($par + comp$)	le comité de suivi ÉVIDEMMENT **intercongolais**
Vocabulaire ouvert		le comité de suivi ÉVIDEMMENT **intercongolais**

Exemple 3

Référence		des dizaines de femmes vêtues de **djellabas** bleues
Sans enrichissement		des dizaines de femmes vêtues de DJELLABA BLEU
Vocabulaire fermé	+ unigrammes uniquement	des dizaines de femmes vêtues de DJELLABA BLEU
	+ n-grammes (par)	des dizaines de femmes vêtues de **djellabas** bleues
	+ n-grammes ($par + comp$)	des dizaines de femmes vêtues de **djellabas** bleues
Vocabulaire ouvert		des dizaines de femmes vêtues de **djellabas** LE

TABLE 7.5 – Exemples de groupes de souffles contenant des mots hors vocabulaire (en gras) transcrits à partir de différentes configuration du système avant et après enrichissement du vocabulaire. En majuscules, les erreurs de transcription.

		Développement	Test
Vocabulaire fermé	Sans enrichissement	23,3	22,4
	Oracle	22,7 $(-0,6)$	21,8 $(-0,6)$
	Unigrammes uniquement	23,0 $(-0,3)$	22,2 $(-0,2)$
	N-grammes (par)	**22,8** $(-0,5)$	**22,1** $(-0,3)$
	N-grammes ($par + comp$)	22,9 $(-0,4)$	**22,1** $(-0,3)$
Vocabulaire ouvert	Sans enrichissement	23,4	22,5
	Oracle	22,8 $(-0,6)$	21,9 $(-0,6)$
	Avec enrichissement	**22,9** $(-0,5)$	**22,2** $(-0,3)$

TABLE 7.6 – WER mesurés sur les ensembles de développement et de test pour différentes techniques d'intégration des mots hors vocabulaire dans un modèle de langue pré-existant.

		Développement	Test
Vocabulaire fermé	Sans enrichissement	100	100
	Unigrammes uniquement	50	66
	N-grammes (par)	35	32
	N-grammes ($par + comp$)	33	31
Vocabulaire ouvert	Sans enrichissement	100	100
	Avec enrichissement	**21**	**21**

TABLE 7.7 – WER mesurés uniquement sur les mots communs thématiques hors vocabulaire (WER_{HV}).

7.3 Bilan du chapitre

Ce chapitre nous a permis de présenter des travaux exploratoires que nous avons menés sur l'adaptation thématique du vocabulaire. Après avoir étudié le potentiel des corpora thématiques pour retrouver des mots hors vocabulaire, notre principale contribution réside dans une méthode originale d'apprentissage de n-grammes attachés à de nouveaux mots du système. Nous avons montré que l'intégration de ces n-grammes dans un modèle de langue pré-existant permet une reconnaissance d'environ 70 % des mots hors vocabulaire ajoutés au système. Bien que ce chiffre soit légèrement inférieur à celui reporté en utilisant un modèle de langue à vocabulaire ouvert, notre technique permet une meilleure intégration des nouveaux mots dans le modèle de langue. Beaucoup de travail reste encore à accomplir cependant pour que notre méthode s'intègre dans un processus complètement non supervisé tel que nous le souhaitons.

À notre sens, le nœud du problème tient principalement dans l'étape de découverte des mots à ajouter au système. Celle-ci devrait s'appuyer sur un mécanisme de génération automatique de variantes linguistiques, notamment compositionnelles, et sur un filtrage conjointement fondé sur les notions d'importance thématique et d'adéquation acoustique.

Ensuite, sur le plan de l'intégration de nouveaux mots, nous n'avons pas traité le problème des noms propres tels que les noms de personnes ou de lieu, noms qui représentent pourtant une part importante des mots hors vocabulaire d'un segment. Nous pensons cependant que ceux-ci doivent pouvoir être intégrés efficacement dans un modèle de langue par l'utilisation de classes ouvertes car, contrairement aux mots que nous avons considérés, ces noms partagent des usages syntaxiques communs.

Outre d'autres aménagements tels que la nécessité probable d'améliorer la qualité de nos relations paradigmatiques, nous pensons finalement qu'il serait véritablement intéressant de chercher à reproduire notre méthode sur la base d'un modèle factoriel. L'existence de caractéristiques linguistiques explicites pour chaque mot au sein des n-grammes recensés par le système faciliterait en effet la recherche de n-grammes équivalents.

Conclusion

Bilan de la thèse

L'objectif de cette thèse a été d'étudier l'adaptation thématique des composants linguistiques d'un système de reconnaissance automatique de la parole. Plus particulièrement, nous nous sommes fixé comme axe de travail de rester dans un cadre le moins supervisé possible en évitant l'usage de toute connaissance *a priori*, notamment sur les thèmes rencontrés, et en ayant recours à des techniques de traitement automatique des langues. Par ailleurs, nous nous sommes volontairement restreint à l'emploi de modèles de langue n-grammes car ces derniers sont les plus performants pour la tâche de reconnaissance automatique de la parole, notamment en raison de leur intégration aisée dans le processus de décodage, et qu'ils sont ainsi les plus répandus actuellement. Dans ce cadre, nous avons proposé une chaîne de traitements originale, complète et entièrement non supervisée. Celle-ci vise à construire un corpus spécifique à chaque thème rencontré et à effectuer l'adaptation du modèle de langue et du vocabulaire du système de reconnaissance. Chacune de ces étapes a fait l'objet d'une contribution notable.

Partant d'un segment traitant d'un thème unique mais inconnu à l'avance, nous avons proposé de caractériser le thème de ce segment par des mots-clés extraits automatiquement en aménageant le critère *tf-idf* , notamment par la prise en compte des mesures de confiance retournées par le système de transcription, afin de rendre cette étape robuste face aux spécificités de l'oral. En lançant des requêtes dans un moteur de recherche sur Internet à partir de ces mots-clés, nous récupérons un ensemble de pages Web potentiellement liées au segment considéré. Nous avons alors proposé de filtrer thématiquement ces pages en mesurant la similarité thématique du contenu de chaque page avec celui du segment, ce qui produit une amélioration notable de la qualité thématique des corpora. En utilisant ces corpora pour adapter un modèle de langue généraliste, nous montrons que cette chaîne de traitements permet de mieux modéliser le langage utilisé au sein d'un thème puisque, à partir des modèles de langue ainsi adaptés, nous obtenons des diminutions de la perplexité et produisons des gains sur le WER des segments thématiques.

Dans un deuxième temps, nous avons cherché à tirer le meilleur parti des corpora thématiques générés automatiquement afin d'adapter au mieux le modèle de langue généraliste de notre système. Notre contribution principale à cet égard a consisté à proposer d'utiliser le cadre d'adaptation par minimum d'information discriminante pour contraindre la réestimation d'un modèle de langue généraliste en s'appuyant sur un ensemble de termes

simples propres au thème considéré et extraits automatiquement. Dans ce cadre, nous avons notamment montré que très peu de termes simples peuvent suffire pour atteindre voire dépasser les gains sur le WER produits par des méthodes classiques telles que l'interpolation linéaire ou l'*unigram rescaling*. Par ailleurs, nous avons montré que l'utilisation de termes complexes est une tâche plus ardue étant donné leur grande variabilité linguistique et leur longueur parfois trop grande pour pouvoir être modélisés par des séquences n-grammes. Enfin, au cours de cette étude, nous avons également souligné que l'intérêt d'utiliser un modèle de langue adapté est affaibli par la stratégie de décodage mise œuvre au sein du processus de transcription pour garantir un temps de calcul acceptable. Il ressort ainsi qu'il est préférable d'intégrer un modèle de langue adapté au plus tôt dans le processus de transcription mais que cette intégration devrait, dans l'idéal, s'accompagner de traitements supplémentaires, notamment au niveau de l'algorithme de création des graphes de mots.

Dans une dernière partie plus prospective, nous avons étudié la question de l'adaptation thématique du vocabulaire d'un système généraliste en portant une attention particulière aux mots hors vocabulaire liés au thème d'un segment, hormis les noms propres. Dans ce contexte, nous avons souligné la difficulté à sélectionner des mots pertinents à ajouter au vocabulaire. Nous avons notamment insisté sur le besoin de développer un processus de filtrage performant d'une liste de mots-candidats en s'appuyant tant sur un critère phonétique que sur un critère de pertinence thématique. Par ailleurs, nous avons également proposé une méthode originale d'intégration de nouveaux mots dans un modèle de langue pré-existant. Par le recours à des relations paradigmatiques entre mots, nous déterminons automatiquement les n-grammes au sein desquels les nouveaux mots devraient en toute logique apparaître et ajoutons ces séquences dans un modèle de langue n-gramme classique. Cette méthode permet d'obtenir des performances comparables à l'utilisation d'un modèle de langue à vocabulaire ouvert qui, lui, ne permet pas de préciser sémantiquement l'emploi des nouveaux mots du vocabulaire.

Perspectives

Dans la continuité de notre thèse, diverses perspectives de travail se dégagent. Nous dressons dans cette dernière section celles d'entre elles qui nous semblent les plus pertinentes pour les facettes techniques, théoriques et applicatives de notre travail.

Tout d'abord, des améliorations pourraient être apportées aux différentes étapes du processus d'adaptation que nous avons présenté. Dans un premier temps, l'extraction de mots-clés utilisée pour la construction de corpora thématiques devrait être rendue plus robuste. Pour cela, nous avons déjà souligné le fait que l'utilisation de relations sémantiques entre mots devrait permettre d'assurer une meilleure cohérence des mots-clés et d'éviter l'émergence d'erreurs de transcription parmi ceux-ci. Dans cette perpective, le calcul de mesures de confiance plus fiables devrait lui aussi être étudié. Dans l'état actuel du système IRENE, ces mesures ne tiennent compte que des scores fournis par nos modèles acoustiques et linguistiques ainsi que d'un post-traitement fondé sur l'utilisation de la morphosyntaxe. Des techniques récentes permettent pourtant de s'appuyer sur d'autres caractéristiques telles que le repérage d'entités nommées ou le niveau de repli atteint par le modèle de langue pour la transcription d'un mot (Fayolle *et al.*, 2010). Dans un même ordre d'idée, l'adaptation thématique d'un modèle de langue n-gramme devrait également s'appuyer sur une meilleure mise en lumière des spécificités linguistiques d'un thème. Ceci passe par l'uti-

lisation de techniques plus sophistiquées et plus complètes d'acquisition de termes simples et de termes complexes. Dans un deuxième temps, les freins à l'intégration des connaissances linguistiques au sein du modèle de langue devraient être levés. Cela devrait passer par l'étude de méthodes efficaces pour ajouter de nouveaux n-grammes à un modèle préexistant. Bien que nous ayons en partie abordé ce problème au chapitre 7, sa résolution complète nécessite encore des réflexions autour de la gestion du lissage et des coefficients de repli. Il serait également bon d'intégrer des caractéristiques morphologiques et sémantiques des mots directement dans le modèle de langue. En cela, l'utilisation de modèles factoriels représente, à notre sens, un axe de travail primordial. Dans un dernier temps, l'adaptation du vocabulaire devrait être complétée par la mise en œuvre d'une méthode de sélection des mots à ajouter à un système. Comme nous l'avons déjà abordé, cette tâche est particulièrement difficile étant donné le nombre de mots-candidats qui émanent d'un corpus thématique extrait du Web.

Par ailleurs, si l'on revient vers des aspects plus théoriques, le problème de l'adaptation thématique présente deux difficultés principales que nous avons cherché à concilier : la formalisation de l'emploi du langage au sein d'un thème donné et l'intégration de ces connaissances au sein de n-grammes. À notre sens toutefois, après l'étude approfondie que nous venons de présenter dans ce manuscrit, la résolution conjointe de ces deux problèmes est en partie utopique. En effet, la description du langage au sein d'un thème nécessite de faire abstraction des possibilités de variantes linguistiques qu'offre le langage naturel et passe en cela par la mise en évidence de relations entre mots non contigus dans un énoncé. Or, une modélisation linguistique fondée sur des n-grammes demande de savoir lister explicitement et de la manière la plus exhaustive possible des séquences de quelques mots contigus. Comme il s'avère irréaliste de lister l'ensemble des réalisations linguistiques spécifiques à un thème donné, il nous semble donc que l'usage de modèles n-grammes seuls n'est pas souhaitable. Il serait véritablement intéressant d'étudier l'introduction de modèles additionnels permettant la prise en compte de relations distantes entre mots au sein du processus de transcription. Les modèles à amorces présentent en cela une base très intéressante à partir de laquelle il faudrait chercher à introduire la notion d'adaptabilité et de spécialisation à chaque thème rencontré par le système de transcription. Pour rester compétitif en terme de temps de calcul avec un modèle n-gramme classique, il reste néanmoins nécessaire de réfléchir également à un moyen d'intégrer ces modèles à amorces dans un algorithme de décodage efficace.

Finalement, notre méthode d'adaptation devrait pouvoir profiter à des applications liées au traitement automatique de documents multimédias. Notamment, l'amélioration du taux de reconnaissance des termes d'un document devrait être mise à profit dans le cadre d'un système d'indexation de documents audio. Toutefois, des réflexions sont à apporter à notre méthode pour transposer celle-ci à des quantités de documents propres à ce type d'application, par exemple traiter quotidiennement toutes les chaînes de télévisions françaises. Plus particulièrement, il faudrait chercher des solutions pour minimiser le coût d'une adaptation thématique. Essentiellement, il serait bon de savoir diagnostiquer si, oui ou non, une adaptation thématique nécessite d'être effectuée pour les besoins d'un document donné. En effet, il se peut très bien qu'un document à transcrire ou à indexer ne soit pas véritablement marqué thématiquement, auquel cas l'emploi des modèle de langue et vocabulaire initiaux suffisent. Pour cela, il serait intéressant de chercher à transposer le concept de mesure de confiance sur les mots à une mesure de confiance à l'échelle d'un segment complet. Outre les indices acoustiques et linguistiques fournis par le système lui-

même, cette dernière mesure pourrait s'appuyer sur des scores *tf-idf* tels que ceux que nous avons utilisés pour caractériser le thème de chaque segment, en ce sens que l'absence de mots très discriminants pourrait signaler l'absence d'un véritable thème. Cependant, ce phénomène pourrait également traduire la présence de nombreux mots hors vocabulaire. À moins d'avoir un accès privilégié à une base d'index du Web[42], l'optimisation du processus d'adaptation passe également par une réduction du nombre de fois où un corpus thématique doit être construit à partir d'Internet. Pour cela, il serait intéressant de chercher à opérer des regroupements entre documents multimédia traitant d'un même thème. Ce regroupement se justifierait particulièrement dans certains cadres d'applications, comme le traitement de flux télévisuels d'actualités, au sein desquels l'information se répète beaucoup d'une chaîne à l'autre. Nous avons présenté au sein de notre travail une mesure de similarité entre documents qui pourrait être utilisée à cet effet. De la même manière, certains documents reprennent des traits thématiques de documents déjà rencontrés par le passé comme, par exemple, lorsqu'un sujet d'actualité se développe de jour en jour mais garde finalement toujours le même débat de fond. Ainsi, il serait pertinent de conserver les documents déjà récupérés lors d'adaptations passées pour éviter de dispendieuses recherches à outrance sur Internet. Enfin, le coût d'une adaptation thématique tire également sa source de l'exécution à deux reprises d'un processus de transcription complet – l'un pour caractériser le thème d'un document, l'autre pour le décodage après adaptation du système. Ainsi, une piste de travail intéressante serait de faire remonter la tâche de caractérisation d'un thème le plus en amont possible dans le processus de transcription. Différents travaux issus du domaine de l'indexation vont d'ailleurs déjà dans ce sens en s'attaquant au traitement de graphes de mots comme support de l'information en sus de transcriptions automatiques (Chelba et Acero, 2005).

[42] Comme c'est le cas pour des entreprises comme *Google*, *Yahoo !* ou *Exalead*.

Annexe A

Démonstrations et détails mathématiques

A.1 Somme à 1 dans un modèle de langage avec *backoff*

A.1.1 Somme des probabilités conditionnelles

Soit la définition du coefficient de repli pour un historique h :

$$\beta(h) \quad = \quad \frac{1 - \sum_{hw \in E} P^*[w|h]}{1 - \sum_{hw \in E} P^*[w|h^-]} \ . \tag{A.1}$$

L'historique h étant fixé, on peut exprimer $\sum_{hw \in E} P^*[w|h^-]$ comme $\sum_{w \in E_h} P^*[w|h^-]$. Or, par définition,

$$\sum_{w \in E_h} P^*[w|h^-] + \sum_{w \notin E_h} P^*[w|h^-] = 1 \ . \tag{A.2}$$

D'où, par analogie,

$$\beta(h) \quad = \quad \frac{1 - \sum_{hw \in E} P^*[w|h]}{\sum_{hw \notin E} P^*[w|h^-]} \ . \tag{A.3}$$

On en déduit facilement que

$$\sum_{hw \in E} P^*[w|h] + \beta(h) \times \sum_{hw \notin E} P^*[w|h^-] \quad = \quad 1 \tag{A.4}$$

$$\sum_{w \in V} P[w|h] \quad = \quad 1 \ . \square \tag{A.5}$$

A.1.2 Somme des probabilités jointes

Par définition, on a $\sum_{w \in V} P[w] = 1$.
Supposons que, pour tout historique h de longueur $n-1$, on a $\sum_{h \in V^{n-1}} P[h] = 1$.

À l'ordre n, on a :

$$\sum_{hw \in V^n} P[hw] = \sum_{hw \in V^n} P[w|h] \times P[h] \tag{A.6}$$

$$= \sum_{w \in V} P[w|h] \times \sum_{h \in V^{n-1}} P[h] \; . \tag{A.7}$$

D'après (A.5), nous simplifions par

$$\sum_{hw \in V^n} P[hw] = \sum_{h \in V^{n-1}} P[h] = 1 \; . \tag{A.8}$$

Par récurrence, on a donc finalement :

$$\sum_{hw \in V^n} P[hw] = 1, \forall n \in \mathbb{N}^* \; .\square \tag{A.9}$$

A.2 La perplexité comme une fonction de l'entropie croisée

La perplexité peut s'exprimer dans le cadre de la théorie de l'information comme une fonction de l'entropie croisée (Cover et Thomas, 1991).

Si l'on considère une variable aléatoire X, symbolisant la source d'information, et une distribution de probabilité P sur l'ensemble $\mathcal{X}$ des valeurs possibles de X, l'entropie de X se calcule par :

$$H(X) = -\sum_{x \in \mathcal{X}} P[x] \times \log_2 P[x] \; . \tag{A.10}$$

Dans ce cadre, supposant la distribution P sur l'ensemble $\mathcal{L}$ des séquences de mots W dans la langue, la qualité d'un modèle de distribution P_M pour prédire les séquences de $\mathcal{L}$ peut se poser comme l'entropie croisée entre P et P_M :

$$H(P, P_M) = -\sum_{W \in \mathcal{L}} P[W] \times \log_2 P_M[W] \; . \tag{A.11}$$

Comme $\mathcal{L}$ est infini et P est inconnue, cette formule ne peut être directement appliquée. Comme le langage naturel décrit un processus stochastique stationnaire et ergodique (Cover et Thomas, 1991), $\mathcal{L}$ peut toutefois être remplacé par un texte $T = w_1 ... w_n$ censé être représentatif de $\mathcal{L}$. On a alors la relation :

$$\lim_{n \to +\infty} H(P_T, P_M) = H(P, P_M) \; . \tag{A.12}$$

Ainsi, pour n suffisamment grand, on peut approcher la formule (A.11) par :

$$H(P, P_M) \approx -\log_2 P_M[T]$$

$$= -\log_2 \prod_{w_i \in T} P_M[w_i|h_i]$$

$$= -\sum_{w_i \in T} \log_2 P_M[w_i|h_i] \; . \tag{A.13}$$

On définit alors le taux d'entropie croisée par mot d'un texte T par rapport au modèle s'écrit comme :

$$\widetilde{H}(P_T, P_M) = -\frac{1}{n} \times \sum_{w_i \in T} \log_2 P_M[w_i|h_i] \ , \tag{A.14}$$

Notons que la base 2 du logarithme permet d'interpréter $L(T|M)$ comme le nombre moyen de bits utilisés par M pour encoder chaque mot de T.

Par analogie avec les formules fondées sur la log-vraissemblance, la perplexité de T sachant M peut alors s'écrire :

$$\mathrm{PPL}_M(T) = 2^{\widetilde{H}(P_T, P_M)} \ . \tag{A.15}$$

A.3 Facteur de mise à l'échelle des *n*-grammes non contraints

L'expression générale du facteur de mise à l'échelle d'un n-gramme hw est :

$$\alpha(hw) = \prod_{i=1}^{K} \alpha_i(hw)^{\frac{1}{\chi_{hw}}}, \tag{A.16}$$

$$\text{avec } \alpha_i(hw) = \left(\frac{K_i}{\langle f_i, P_B \rangle} \right)^{f_i(hw)} \ , \tag{A.17}$$

où χ_{hw} est le nombre de caractéristiques, c'est-à-dire de contraintes, auxquelles appartient hw et les f_i sont des fonctions indicatrices propres à chaque caractéristique.

Si le n-gramme hw n'appartient à aucune contrainte, on peut poser que :

$$f_i(hw) = 0, \forall i \ . \tag{A.18}$$

Donc,

$$\alpha_i(hw) = 1, \forall i \ . \tag{A.19}$$

Or, comme $\lim_{x \to 0} 1^{\frac{1}{x}} = 1$, on a

$$\lim_{\chi_{hw} \to 0} \alpha(hw) = \lim_{\chi_{hw} \to 0} \prod_{i=1}^{K} \alpha_i(hw)^{\frac{1}{\chi_{hw}}} = 1 \ .\square \tag{A.20}$$

Annexe B

Exemples de segments thématiques

Dans cette annexe, nous présentons deux segments thématiques avec leur transcriptions manuelle et automatique (section 1) pour lesquels nous donnons les résultats lors l'extraction des mots-clés (section 2).

Le premier exemple représente un cas où tout se passe bien. Il s'agit d'un segment sur la mise en place de système de radar automatique le long des routes françaises. Il comporte peu de noms propres et d'erreurs de transcriptions.

Le second illustre un cas où l'adaptation thématique échoue. Ce segment est un long reportage effectué au sein d'un village, nommé Lieudieu, ayant massivement voté pour Jean-Marie Le Pen aux élections présidentielles de 2002. Une majeure partie de la parole y relève de la parole spontanée puisque le reportage est principalement constitués de témoignages d'habitants interviewés dans la rue. Par ailleurs, ce segment contient beaucoup de noms propres qui reviennent fréquemment dans le discours. Il en découle une transcription automatique de relativement mauvaise qualité avec une faible richesse lexicale.

B.1 Transcriptions

Exemple 1

Transcription manuelle	Transcription automatique
souriez si vous prenez la route c' est la journée de la courtoisie au volant à l' occasion du week-end pascal qui s' annonce très chargé rouge aujourd' hui et demain et lundi dans le sens des retours souriez mais surtout soyez vigilants d' autant que les policiers eux sont de plus en plus vigilants grâce à de nouveaux radars lancés dans le cadre de la lutte contre l' insécurité routière romain auzui fini le temps des barbecues ces gros radars à lunettes qu' on peut voir sur les autoroutes bientôt tout sera automatisé à la place du système magnétique qui vous flashe quand vous êtes en excès de vitesse on aura droit à un radar numérique le but gagner en main-d'oeuvre plus besoin d' un gendarme qui prend la photo mais surtout économiser du temps puisque le numérique qui transmet l' image quasiment en direct remplacera le développement assez long des photos par conséquent les contrôles seront plus rapides et plus nombreux le fonctionnement est simple le radar sera installé sur une caméra numérique donc quand vous serez en excès de vitesse celle-ci déclenchera automatiquement une impulsion et la photo sera transmise vers un ordinateur connecté à une sorte de grand fichier central des cartes grises à ce moment-là on pourra connaître le numéro de votre plaque d' immatriculation et votre amende sera immédiatement imprimée puis envoyée ces nouveaux radars commenceront à être mis en place dès la fin de l' année avec l' installation d' une centaine d' entre eux l' objectif c' est qu' il y en ait mille en deux mille quatre essentiellement sur les autoroutes et les routes nationales	sauriez si vous prenez la route c' est la journée de la courtoisie au volant l' occasion du week-end pascal qui s' **annoncent recharger** rouge aujourd'hui et demain et lundi dans le sens des retours **selon les** mais surtout soyez vigilants d' autant que les policiers **ne** sont de plus en plus vigilants grâce à de nouveaux radars lancés dans le cadre de la lutte contre l' insécurité routière **en main aux oui** fini le temps des barbecue **ses** gros **radar** à lunettes qu' on peut voir sur les autoroutes **et** bientôt tout sera automatisé à la place du système magnétique qui vous **flashes** quand vous êtes en excès de vitesse on aura droit un radar numérique le but de gagner en main-d'oeuvre plus besoin d' un gendarme qui prend la photo mais surtout économiser du temps puisque le numérique qui transmet les images quasiment en direct remplacera le développement assez long des photos par conséquent les contrôles seront plus **rapide** et plus nombreux le fonctionnement est simple le radar sera installé sur une caméra numérique **non** quand vous serez en excès de vitesse celle ci déclenchera automatiquement une impulsion et la photo **soit transmis** vers un ordinateur **se connecter** à une sorte de grand fichier central des cartes grises **a souvent** là on pourra connaître le numéro de votre **pragmatique nation** et votre **monde** sera immédiatement **imprimer** puis **envoyées** ces nouveaux radars commenceront à être mis en place dès la fin de l' année avec l' installation d' une centaine d' entre eux l' objectif est **étudiant les mines** en deux mille quatre essentiellement sur les autoroutes et routes nationales

TABLE B.1 – Transcriptions manuelle et automatique du premier segment. En gras, les erreurs de transcription.

Exemple 2

Transcription manuelle	Transcription automatique
vingt et un avril deux mille deux un an après à quarante huit heures près euh france inter termine ce matin la série de reportages consacrés au séisme politique du premier tour de l' élection présidentielle aujourd' hui nous vous proposons une photographie d' un village qui a massivement voté front national pas dans le nord ou l' est de la france où jean-marie le pen a dans l' ensemble réalisé de bons scores non dans un village de l' isère à lieudieu dans une région où les scores du front national étaient légèrement supérieurs à la moyenne nationale sauf ce vingt et un avril à lieudieu jean-marie le pen a décroché trente six pour cent des voix au premier tour et même le second tour a été serré cinquante quatre pour cent pour jacques chirac quarante six pour cent pour jean-marie le pen reportage à lieudieu vanessa descouraux deux cent soixante et onze habitants c' est calme lieudieu il y a l' école à côté là c' est la mairie et là c' est la salle de location nous on l' a louée la première pour notre anniversaire de mariage et des étangs tout autour du village des fermes perdues au bout de sentiers improbables et des maisons en construction car lieudieu n' est qu' à une soixantaine de kilomètres de lyon et la commune est devenue le refuge des citadins qui quittent la ville ou la banlieue des gars qui s' échappent de la banlieue c' est pas là qu' ils vont changer d' idées hein moi je suis de vaulx-en-velin ma femme elle est de saint-priest [...] les gens disent oh oui bon on est peut-être allé un peu fort mais malgré tout bon ça prouve que les gens euh demandaient un peu plus de de sévérité de discipline ce sont peut-être les premiers maintenant qui rouspètent parce que il y a des des lois qui sont peut-être un peu plus strictes le maire actuel ne souhaite plus évoquer ce qu' il appelle un accident enfin j' espère que c' était un accident précise -t-il	le vingt-et-un avril deux mille deux un an après à quarante-huit heures près euh france-inter termine ce matin la série de reportages consacrés au séisme politique du premier tour de l' élection présidentielle aujourd'hui nous vous proposons une photographie d' un village qui a massivement voté front national **pardon** nord **où** l' est de la france jean-marie le pen **à** l' ensemble réalisé de bons scores dans dans un village de l' isère **est un lieu dieu** dans une région où les scores du front national **était** légèrement **supérieur** à la moyenne nationale sauf que ce vingt-et-un avril **elle a lieu du** jean-marie le pen a décroché trente-six pour cent des voix au premier tour **de** même le second tour a été serré cinquante-quatre pour cent pour jacques chirac quarante-six pour cent pour jean-marie le pen reportage **a lieu dieu** vanessa descouraux deux cent soixante et onze habitants **se** calme **lieu dieu il faut à la fois** et là c' est la mairie **hélas** la salle de **ceux** location **mais** on l' a loué la première année **bon** anniversaire de mariage des **attentats** tout autour du village des fermes perdues au bout de sentiers improbables et des maisons en construction car **lieu dieu et** une soixantaine de kilomètres de lyon et la commune est **devenu** le refuge des citadins qui quittent la ville **où** la banlieue **il est envisageable** de la banlieue **séparer qui** vont changer d' **idée** moi je **vends un peu** ma femme **et** saint-priest [...] **ça se rendre dans l' ordre depuis un an avant d' être aller de soi mais dans la tombe on se retrouve que de** gens **sont demandeurs de** plus de **ce dossier et** de discipline **sans fond du problème à passer aux pas parce que euh des** des lois qui **sentent un** peu plus **près** le maire actuel ne **soit** plus **évoqué** ce qu' il appelle un accident enfin j' espère que c' était un accident **précisent ils**

TABLE B.2 – Transcriptions manuelle et automatique du second segment. En gras, les erreurs de transcription.

B.2 Mots-clés

Exemple 1

$\sigma(\ell)$	Mot le plus fréquent	Classe d'équivalence ℓ
0,88930	radar	{radar, radars}
0,45378	numérique	{numérique}
0,35375	photo	{photo, photos}
0,31960	autoroute	{autoroutes}
0,31902	automatisé	{automatisé}
0,30146	flashes	{flashes}
0,29887	excès	{excès}
0,26791	vitesse	{vitesse}
0,26694	barbecue	{barbecue}
0,25378	vigilants	{vigilants}

TABLE B.3 – Liste des 10 mots-clés pour l'exemple 1.

Exemple 2

$\sigma(\ell)$	Mot le plus fréquent	Classe d'équivalence ℓ
0,96991	pen	{pen}
0,46765	dieu	{dieu}
0,39670	descouraux	{descouraux}
0,38914	lieu	{lieu, lieux}
0,37984	réalisé	{réalisé}
0,36785	perturbé	{perturbée}
0,35911	saint-priest	{saint-priest}
0,35562	scores	{scores, score}
0,33015	makaïla	{makaïla}
0,30230	village	{village}
0,25402	front	{front}

TABLE B.4 – Liste des 10 mots-clés pour l'exemple 2.

Annexe C

Nettoyage des pages Web

L'étude des premiers corpora récupérés a révélé le caractère bruité et peu utilisable de certaines pages qui contiennent pourtant des textes thématiquement très proches de ce qui est recherché. En effet, dans la majorité des cas, une page regroupe un bandeau de titre, un menu, des mentions légales, des résumés d'articles récents, des publicités, des commentaires et, finalement, un article central. Cependant, comme les données récupérées se présentent sous la forme de code HTML, aucune image, ou autre entité physiquement séparée de la page, n'est rapatriée. Cette annexe présente ainsi la technique que nous avons développé pour nettoyer la source d'une page Web de manière à en extraire les articles centraux et les commentaires qui leur sont parfois attachés. Cette méthode s'appuie sur deux passes : l'une servant à éliminer les éléments clairement identifiés comme n'étant pas un texte bien construit, l'autre visant à ne conserver que la zone de texte la plus dense dans la page en excluant les passages ayant passé le premier filtre qui ne relèvent cependant pas du contenu central de la page (par exemple, des résumés de brèves d'actualités d'une même journée). Nous exposons brièvement le principe de ces deux passes et présentons un exemple de page nettoyée.

La structure d'un code HTML est celle d'un arbre dont les feuilles sont du texte ou des balises vides. Ses noeuds sont simplement des balises. Le but de notre première passe est alors de faire remonter les feuilles qui contiennent des bouts de textes que nous jugeons bien construits. En procédant par un parcours horizontal de l'arbre en partant des feuilles, nous remplaçons chaque noeud par la concaténation du texte qui a été conservé dans chacune de ses branches. Au final, lorsque nous arrivons à la racine, il nous reste un texte utilisable pour l'apprentissage d'un modèle de langue et d'où toutes les balises HTML sont supprimées. Le principal critère qui nous sert à conserver ou écarter une feuille se base sur la fréquence de la ponctuation qu'elle contient. Par l'observation de quelques pages prises au hasard, nous avons constaté que, dans un texte construit, il y a en moyenne un signe de ponctuation pour dix mots. À une marge de tolérance près, nous éliminons alors les feuilles dont le texte est trop peu ponctué. De même, si des signes de ponctuation sont rencontrés trop fréquemment, la feuille n'est pas conservée. De plus, nous ne gardons pas les textes trop courts car ceux-ci sont généralement des références vers des actualités récentes. Nous avons également déterminé des intervalles de confiance pour la fréquence des caractères spéciaux rencontrés (par exemple `|`, `+`, `*`...) et pour la longueur moyenne des phrases que nous manipulons. Enfin, il a été fait en sorte, dans notre algorithme, que la décision d'éliminer une feuille se prenne à partir des résultats sur l'ensemble de nos critères.

Ian Huntley, un ancien gardien d'école comparaît à Londres pour les meurtres de Holly Wells et Jessica Chapman, deux fillettes de dix ans dont la mort en août 2002 avait suscité une vive émotion en Grande-Bretagne. Inculpé de double meurtre, cet individu de 29 ans, qui a toujours clamé son innocence, affronte ainsi la justice plus d'un an après les faits. Le 4 août 2002, les deux écolières du lycée de Soham, petite bourgade sans histoire de l'est de l'Angleterre, avaient disparu, déclenchant la plus importante chasse à l'homme jamais organisée en Grande-Bretagne, avec la mobilisation de près de 400 policiers. Leurs corps sans vie avaient été retrouvés 13 jours plus tard dans un bois des environs. Les jeunes amies partageaient la même passion pour le football et le club de Manchester United. En hommage, une minute de silence est d'ailleurs toujours observée au début de tous les matches de football de Grande-Bretagne. Dans le box des accusés, se tiendra également l'ancienne petite amie de I. Huntley, Maxine Carr, assistante scolaire dans l'école de Holly et Jessica. La jeune femme de 26 ans, est, quant à elle, inculpée de complicité de meurtre et de tentative d'entrave à la justice pour avoir menti aux enquêteurs de la police. Elle a également toujours nié les faits. Le procès, qui durera au moins deux mois, promet d'être l'un des plus médiatiques de ces dernières années. S'il est reconnu coupable des deux meurtres, Ian Huntley encourt une peine de 50 ans de prison.

TABLE C.1 – Texte extrait à partir de la page présentée en figure C.1.

Ainsi, une feuille n'est pas directement éliminée dès lors qu'elle ne répond pas à l'un de nos critères d'exigence. Ce choix est dû au fait que la diversité des structures des pages Web est telle qu'aucune généralité ne peut en être faite avec certitude.

À l'isue de cette première passe, nous récupérons une séquence de paragraphes séparés par des sauts de ligne plus ou moins nombreux. Le nombre de ces sauts de lignes révèle à quel point deux portions consécutives de textes étaient éloignées dans la page Web. Nous définissons alors la position d'un paragraphe dans cette séquence comme une variable aléatoire. Nous calculons la valeur moyenne et la variance de cette variable en pondérant chaque position par le nombre de mots présents à celle-ci. Ces deux valeurs nous permettent de définir d'une loi normale centrée et d'exclure les portions de texte présentes à des positions trop peu probables. De cette manière, nous ne gardons que la zone la plus dense de la page et supprimons tout texte trop éloigné de cette zone.

La figure C.1 présente une page récupérée lors de l'une de nos adaptations. Au sein de celle-ci, nous avons encadré en rouge la partie que nous voudrions conserver. Nous notons que, en dehors de l'article principal, beaucoup de mots isolés sont présents sur la page. Si nous éliminions simplement toutes les balises HTML et que nous concaténions tous les mots de la page, nous obtiendrions un texte dont certains passages seraient syntaxiquement incorrects. Un modèle de langue appris sur cette base serait alors d'une qualité mitigée. Comme le montre la table C.1, notre technique permet d'extraire exactement le texte qui nous intéresse. Notons toutefois que, dans certains cas, notre méthode tend à éliminer trop de texte. C'est notamment le cas lorsque la page ne contient pas beaucoup de texte et que la seconde passe de notre méthode ne parvient pas clairement à déterminer la zone centrale du contenu. Toutefois, cette inconvénient est à pondérer par le fait que, de toute manière, ce type de page ne nous intéresse que peu pour estimer des probabilités n-grammes.

FIGURE C.1 – Exemple de page Web à nettoyer.

Annexe D

Affinement des terminologies de termes simples

Lors de nos premiers sur l'extraction de termes simples, nous avons directement repris la technique de calcul des scores *tf-idf* que nous avions déjà utilisées pour la création automatique de corpora thématiques. Il apparait toutefois qu'un certain bruit persiste dans les terminologies en accordant une importance trop importante à certains mots détâchés du thème et en sous-estimant parfois au contraire d'autres termes vraiment spécifiques. Cette annexe a pour but de montrer que ces problèmes peuvent être traités en restant dans le même cadre que celui proposé durant nos expériences sur l'adaptation d'un modèle de langue.

Lors d'études ultérieures, nous avons cherché à corriger ce problème en modifiant le score S' de chaque lemme de manière) prendre en compte des relations paradigmatiques. L'objectif cette prise en compte est de favoriser les mots qui partagent habituellement un contexte d'apparition similaire. Étant donné deux lemmes ℓ_1 et ℓ_2 supposés partager une relation paradigmatique, nous faisons en sorte que l'apparition de ℓ_1 contribue à augmenter la fréquence d'apparition de ℓ_2. Pour cela, nous calculons la fréquence d'un lemme par la formule :

$$freq(\ell, d) = \frac{N(\ell, d)}{\sum_{x \in d} N(x, d)} \tag{D.1}$$

$$\text{avec } N(\ell, d) = |\ell|_d + \sum_{\ell \mathcal{R} m} \nu \times r(\ell, m), \tag{D.2}$$

où $|\ell|_d$ est le nombre d'occurrences du lemme ℓ dans le document d, $\ell \mathcal{R} m$ signifie que ℓ partage une relation paradigmatique avec m et que $r(\ell, m)$ est le score compris entre 0 et 1 de cette relation. Le paramètre ν permet quant à lui de pondérer l'importance des relations paradigmatiques. La table D.1 montre l'impact de ces relations pour $\nu = 0, 2$ sur le corpus que celui utilisés pour générer les termes simples de la table 6.4. On note ainsi que des mots comme France, pays ou homme ont disparu de la liste des 100 premiers termes sous l'effet des relations paradigmatiques.

Par ailleurs, nous avons voulu expérimenter le calcul d'un autre score pour la fonction *idf* d'un lemme afin de favoriser davantage le caractère rare des mots dans un contexte général par rapport à leur fréquence dans un corpus thématique. Cette volonté vise à faire

ressortir des mots qui, de part leur rareté en général, sont censés être les plus mal estimés par un modèle de langue généraliste. Ainsi, la table D.2 présente les résultats obtenus lorsque la valeur de la fonction *idf* est mise au carré lors du calcul du score S' avec prise en compte des relations paradigmatiques. On remarque parmi ces resultats que la modification a bien l'effet escompté en faisant ressortir des termes quel que équité ou ostentatoire.

laïcités	femme	islamiques	juif	croyance
laïcité	loi	islamique	fondamentale	républiques
musulmans	lois	maternelle	fondamentaux	république
musulman	non-discrimination	culte	fondamental	sarkozy
musulmanes	non-discriminations	cultes	fondamentales	universalité
musulmane	école	communautarismes	signe	universalités
voile	écoles	communautarisme	signes	chrétien
voiles	liberté	égalité	séparation	chrétiens
islam	libertés	égalités	séparations	débat
islams	internationale	confessionnelles	christianisme	débats
religieuses	ostensible	confessionnels	christianismes	tchadors
religieuse	ostensibles	confessionnel	tolérance	tchador
religieux	chrétiennes	confessionnelle	tolérances	droits
religion	chrétien	confessions	voilés	droit
religions	chrétiens	confession	voilé	judéo-chrétienne
laïques	chrétienne	principes	voilée	judéo-chrétiennes
laïque	croyant	principe	voilées	judéo-chrétiens
foulards	croyants	juive	neutralité	judéo-chrétien
foulard	ports	juives	neutralités	respect
femmes	port	juifs	croyances	respects

TABLE D.1 – Liste des 100 mots ayant les meilleures scores après intégration des relation paradigmatiques pour un corpus de 200 pages construits à partir d'un segment traitant de la laïcité et du port du voile.

laïcités	judéo-chrétien	égale[43]	christianismes	ostentatoires
laïcité	religieuses	monothéisme	chrétientés	tolérance
non-discrimination	religieuse	monothéismes	chrétienté	tolérances
non-discriminations	religieux	immigrés	universalité	ports
voile	foulards	immigré	universalités	port
voiles	foulard	tchadors	croyances	laïcs
musulmans	ostensible	tchador	croyance	laïc
musulman	ostensibles	confessions	intangible	islamiques
musulmanes	rendu	confession	intangibles	islamique
musulmane	islamiste	culte	égalité	israélites
maternelle	islamistes	cultes	égalités	israélite
islam	communautarismes	animiste	indivisibles	équité
islams	communautarisme	animistes	indivisible	équités
laïques	croyant	coreligionnaires	neutralité	séparation
laïque	croyants	coreligionnaire	neutralités	séparations
religion	confessionnelles	chrétiennes	judaïsme	etat-providence
religions	confessionnels	chrétien	judaïsmes	subsidiarité
judéo-chrétienne	confessionnel	chrétiens	états-unis	subsidiarités
judéo-chrétiennes	confessionnelle	chrétienne	unicité	école
judéo-chrétiens	égales[43]	christianisme	ostentatoire	écoles

TABLE D.2 – Liste des 100 mots ayant les meilleures scores après intégration des relation paradigmatiques et accentuation du poids de la fonction *idf*.

[41] Il s'agit du nom commun égale et non de l'adjectif.

Bibliographie

Allauzen A. (2003). *Modélisation linguistique pour l'indexation automatique de documents audiovisuels.* Thèse de doctorat, Université Paris 11, France.

Allauzen A. et Gauvain J.-L. (2005a). « Diachronic Vocabulary Adaptation for Broadcast News Transcription ». *Dans Proc. of Intl Conf. on Speech and Language Technology (Interspeech)*, pages 1305–1308, Lisbon, Portugal.

Allauzen A. et Gauvain J.-L. (2005b). « Open Vocabulary ASR for Audiovisual Document Indexation ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1013–1016, Philadelphia, Pennsylvania, USA.

Ariki Y., Shigemori T., Kaneko T., Ogata J. et Fujimoto M. (2003). « Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model ». *Dans Proc. of the European Conf. on Speech Communication and Technology (Eurospeech)*, pages 1453–1456.

Arisoy E. et Saraçlar M. (2009). « Lattice Extension and Vocabulary Adaptation for Turkish LVCSR ». *IEEE Trans. on Audio, Speech, and Language Processing*, 17(1):163–173.

Aronowitz H. (2009). « Online Vocabulary Adaptation using Contextual Information and Information Retrieval ». *Dans Proc. the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 1805–1809.

Auzanne C., Garofolo J. S., Fiscus J. G. et Fisher W. M. (2000). « Automatic Language Model Adaptation for Spoken Document Retrieval ». *Dans Proc. of RIAO 2000 Conf. on Content-Based Multimedia Information Access.*

Baroni M., Chantree F., Kilgarriff A. et Sharoff S. (2008). « Cleaneval : a Competition for Cleaning Web Pages ». *Dans Proc. of the Sixth Intl Language Resources and Evaluation (LREC)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Batista F., Amaral R., Trancoso I. et Mamede N. (2008). « Impact of Dynamic Model Adaptation beyond Speech Recognition ». *Dans IEEE Spoken Language Technology Workshop*, pages 277–280.

Bazzi I. (2002). *Modeling Out-Of-Vocabulary Words For Robust Speech Recognition.* Thèse de doctorat, MIT.

Bazzi I., Schwartz R. et Makhoul J. (1999). « An Omnifont Open-Vocabulary OCR System for English and Arabic ». *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(6):495–504.

Bellegarda J. R. (2000). « Large Vocabulary Speech Recognition with Multispan Statistical Language Models ». *IEEE Trans. on Speech and Audio Processing*, 8(1):76–84.

Bellegarda J. R. (2004). « Statistical Language Model Adaptation : Review and Perspectives ». *Speech Communications*, 42:93–108.

Bengio Y., Ducharme R. et Vincent P. (2001). « A neural probabilistic language model ». *Advances in Neural Information Processing Systems*, 13.

Berger A. et Miller R. (1998). « Just-in-Time Language Modelling ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 705–708, Seattle, Washington, USA.

Bertoldi N. et Federico M. (2001). « Lexicon Adaptation for Broadcast News Transcription ». *Dans Proc. of ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, pages 187–190, Sophia Antipolis, France.

Besling S. et Meier H. (1995). « Language Model Speaker Adaptation ». *Dans Proc. of 4th European Conf. on Speech Communication and Technology (Eurospeech)*. ISCA.

Bigi B., Huang Y. et De Mori R. (2004). « Vocabulary and Language Model Adaptation Using Information Retrieval ». *Dans Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP)*, volume 2, pages 1361–1364, Jeju Island, South Korea.

Bilmes J. A. et Kirchhoff K. (2003). « Factored Language Models and Generalized Parallel Backoff ». *Dans Proc. of Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 4–6.

Bimbot F., Pieraccini R., Levin E. et Atal B. (1995). « Variable-Length Sequence Modeling : Multigrams ». *Signal Processing Letters, IEEE*, 2(6):111–113.

Bonafonte A. et Mariño J. B. (1996). « Language Modeling Using X-Grams ». *Dans Proc. of the 4th Intl Conf. on Spoken Language Processing (ICSLP)*, volume 1, pages 394–397, Philadelphia, Pennsylvania, USA.

Boulianne G., Beaumont J.-F., Boisvert M., Brousseau P., Cardinal J. P., Chapdelaine C., Comeau M., Ouellet P. et Osterrath F. (2006). « Computer-Assisted Closed-Captioning of Live TV Broadcasts in French ». *Dans Proc. of the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 1424–1427, Montréal, Canada.

Brown G. et Yule G. (1983). *Discourse analysis*.

Brown P. F., Della Pietra V. J., deSouza P. V., Lai J. C. et Mercer R. L. (1992). « Class-Based N-Gram Models of Natural Language ». *Computational Linguistics*, 18(4):467–480.

Brun A. (2003). *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. Thèse de doctorat, Université Henri Poincaré - Nancy 1, France.

Brun A., Langlois D. et Smaïli K. (2007). « Improving Language Models by Using Distant Information ». *Dans Proc. of the International Symposium on Signal Processing and its Applications (ISSPA)*, pages 1–4, Sharjah United Arab Emirates.

Bulyko I., Ostendorf M., Siu M., Ng T., Stolcke A. et Çetin O. (2007). « Web Resources for Language Modeling in Conversational Speech Recognition ». *ACM Trans. Speech Lang. Process.*, 5(1):1–25.

Bulyko I., Ostendorf M. et Stolcke A. (2003). « Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures ». *Dans Proc. of the 2003 Conf. of the North AmericanChapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, pages 7–9, Morristown, NJ, USA. Association for Computational Linguistics.

Burget L., Schwarz P., Matejka P., Hannemann M., Rastrow A., White C., Khudanpur S., Hermansky H. et Cernocky J. (2008). « Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4081–4084.

Béchet F. (2001). « LIA_PHON : un système complet de phonétisation de textes ». *Traitement Automatique des Langues (TAL)*, 42(1):47–67.

Cabré Castelliví M. T., Estopà Bagot R. et Vivaldi Palatresi J. (2003). *Recent Advances in Computational Terminology*, chapitre Automatic term detection : a review of current systems, pages 53–88. John Benjamins Publishing co.

Callan J. et Connell M. (2001). « Query-based Sampling of Text Databases ». *ACM Trans. Inf. Syst.*, 19(2):97–130.

Cardillo P. S., Clements M. et Miller M. S. (2002). « Phonetic Searching vs. LVCSR : How to Find what you Really Want in Audio Archives ». *Intl Journal of Speech Technology*, 5(1):9–22.

Carlson B. A. (1996). « Unsupervised Topic Clustering of SWITCHBOARD Speech Messages ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 315–318, Atlanta, Georgia, USA.

Chelba C. et Acero A. (2005). « Position Specific Posterior Lattices for Indexing Speech ». *Dans Proc. of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 443–450, Morristown, New Jersey, USA. Association for Computational Linguistics.

Chelba C., Engle D., Jelinek F., Jimenez V., Khudanpur S., Mangu L., Printz H., Ristad E., Rosenfeld R., Stolcke A. et Wu D. (1997). « Structure and Performance of a Dependency Language Model ». *Dans Proc. of the European Conf. on Speech Communication and Technology (Eurospeech)*, pages 2775–2778.

Chelba C. et Jelinek F. (2000). « Structured Language Modeling ». *Computer Speech and Language*, 14(4):283–332.

Chen L., Gauvain J.-L., Lamel L. et Adda G. (2003). « Unsupervised Language Model Adaptation for Broadcast News ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 220–223, Hong Kong, China.

Chen L., Gauvain J.-L., Lamel L. et Adda G. (2004). « Dynamic Language Modeling for Broadcast News ». *Dans Proc. of the Intl Conf. on Spoken Language Processing (ICSLP)*, pages 1281–1284, Jeju Island, South Korea.

Chen L. et Huang T. (1999). « An Improved MAP Method for Language Model Adaptation ». *Dans Proc. of the 6th European Conf. on Speech Communication and Technology (Eurospeech)*, pages 1923–1926, Budapest, Hungary.

Chen S. F. et Goodman J. (1998). « An Empirical Study of Smoothing Techniques for Language Modeling ». Rapport technique, Harvard University, Cambridge, Massachusetts, USA.

Chen S. F., Seymore K. et Rosenfeld R. (1998). « Topic Adaptation for Language Modeling using Unnormalized Exponential Models ». *Dans Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 681–684, Seattle, WA , USA.

Chueh C.-H. et Chien J.-T. (2008). « Reliable Feature Selection for Language Model Adaptation ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech,and Signal Processing (ICASSP)*, pages 5089–5092.

Chung G., Seneff S., Wang C. et Hetherington I. (2004). « A Dynamic Vocabulary Spoken Dialogue Interface ». *Dans Proc. of the Intl Conf. on Speech and Language Technology (ICSLP)*, pages 1457–1460.

Chung T. (2003). « A Corpus Comparison Approach for Terminology Extraction ». *Terminology*, 9(2):221–246.

Church K. W. et Hanks P. (1990). « Word association norms, mutual information, and lexicography ». *Comput. Linguist.*, 16(1):22–29.

Clarkson P. et Robinson T. (1999). « Towards Improved Language Model Evaluation Measures ». *Dans Proc. of the 6th European Conf. on Speech Communication and Technology (Eurospeech)*, volume 5, pages 1927–1930, Budapest, Hungary.

Claveau V. (2009). « Letter-to-Phoneme Conversion by Inference of Rewriting Rules ». *Dans Proc. the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 1299–1302, Brighton, UK.

Claveau V., Sébillot P. et de Beaulieu C. (2004). « Extension de requêtes par lien sémantique nom-verbe acquis sur corpus ». *Dans Actes de la 11ème conférence de Traitement automatique des langues naturelles, TALN'04*, Fês, Maroc.

Claveau V. et Sébillot P. (2004). « Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe ». *Traitement automatique des langues (TAL)*, 45(1):153–182.

Cover T. M. et Thomas J. A. (1991). *Elements of Information Therory.* Wiley-Interscience.

Daille B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Thèse de doctorat, Université de Paris 7.

Daille B. (2002). *Découvertes linguistiques en corpus.* Thèse de doctorat, Université de Nantes.

Darroch J. N. et Ratcliff D. (1972). « Generalized Iterative Scaling for Log-Linear Models ». *Annals of Mathematical Statistics*, 43(5):1470–1480.

De Mori R. et Federico M. (1999). « Language Model Adaptation ». *NATO ASI series. Series F : computer and system sciences*, pages 280–303.

Deerwester S., Dumais S., Furnas G., Landauer T. et Harshman R. (1990). « Indexing by Latent Semantic Analysis ». *Journal of the American society for information science*, 41(6):391–407.

Della Pietra S., Della Pietra V. et Lafferty J. (1997). « Inducing Features of Random Fields ». *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(4):380–393.

Della Pietra S., Della Pietra V., Mercer R. et Roukos S. (1992). « Adaptive language modeling using minimum discriminant estimation ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 633–636, San Francisco, Californie, USA.

Drouin P. (2004). « Spécificités lexicales et acquisition de la terminologie ». *Dans Actes des 7e Journées internationales d'analyse statistique des données textuelles (JADT)*, pages 345–352.

Dubois D. (2006). « Possibility Theory and Statistical Reasoning ». *Computational Statistics & Data Analysis*, 51(1):47–69.

Dumais S. T. (1994). « Latent Semantic Indexing ». *Dans Proc. of the 2nd Text Retrieval Conf. (TREC-2)*, pages 105–116.

Dunning T. (1993). « Accurate Methods for the Statistics of Surprise and Coincidence ». *Comput. Linguist.*, 19(1):61–74.

Dupont P. et Rosenfeld R. (1997). « Lattice Based Language Models ». Rapport technique, Carnegie Mellon University.

Erdoğan H. (2001). « Speech Recognition for a Travel Reservation System ». *Dans Proc. of Intl Conf. on Artificial Intelligence*.

Fayolle J., Moreau F., Raymond C., Gravier G. et Gros P. (2010). « CRF-based Combination of Contextual Features to Improve A Posteriori Word-level Confidence Measures ». *Dans InterSpeech*, Makuari, Japan.

Federico M. (1996). « Bayesian Estimation Methods for N-Gram Language Model Adaptation ». *Dans Proc. of the 4th Intl Conf. on Spoken Language Processing (ICSLP)*, volume 1, pages 240–243, Philadelphia, Pennsylvania, USA.

Federico M. (1999). « Efficient Language Model Adaptation through MDI Estimation ». *Dans Proc. of the 6th European Conf. on Speech, Communication, Technology (Eurospeech)*, volume 4, pages 1583–1586, Budapest, Hungary.

Federico M. (2002). « Language Model Adaptation Through Topic Decomposition and MDI Estimation ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 773–776, Orlando, Florida, USA.

Ferret O. (1998). « How to Thematically Segment Texts by Using Lexical Cohesion ? ». *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Intl Conf. on Computational Linguistics (COLING-ACL)*, volume 2, pages 1481–1483, Montreal, Canada.

Fouqueré C. (2003). « Corpus issus du Web : constitution et analyse informationnelle ». *Revue québécoise de linguistique*, 32:1.

Frantzi K., Ananiadou S. et Mima H. (2000). « Automatic Recognition of Multi-Word Terms : the C-Value/NC-Value Method ». *Intl Journal on Digital Libraries*, 3(2):115–130.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-F. et Gravier G. (2005). « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News ». *Dans Proc. of the 9th European Conf. on Speech Communication and Technology (Eurospeech)*, pages 1149–1152, Lisbon, Portugal.

Gao Q., Lin X. et Wu X. (2006). « Just-in-Time Latent Semantic Adaptation on Language Model for Chinese Speech Recognition using Web Data ». *Dans Proc. of Spoken Language Technology Workshop (SLT)*, pages 50–53.

Geutner P., Finke M. et Scheytt P. (1998a). « Adaptive vocabularies for transcribing multilingual broadcast news ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 925–928 vol.2.

Geutner P., Finke M. et Waibel A. (1998b). « Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation For Transcribing Multlingual Broadcast News ». *Dans Proc. of 5th Intl Conf. on Spoken Language Processing (ICSLP)*, Sidney, Australia.

Ghadessy M. (1995). *Thematic development in English texts*. Pinter Publishers.

Gildea D. et Hofmann T. (1999). « Topic-Based Language Models Using EM ». *Dans Proc. of the 6th European Conf. on Speech Communication and Technology (Eurospeech)*, pages 2167–2170, Budapest, Hungary.

Gillick L. et Cox S. J. (1989). « Some Statistical Issues in the Comparison of Speech Recognition Algorithms ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 532–535, Glasgow, UK.

Good I. J. (1953). « The Population Frequencies of Species and the Estimation of Population Parameters ». *Biometrika*, 40(3-4):237–264.

Gravier G., Bonastre J.-F., Galliano S., Geoffrois E., Mc Tait K. et Choukri K. (2004). « ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques ». *Dans Actes des 25èmes Journées d'Étude sur la Parole (JEP)*, Fes, Morocco.

Gravier G., Yvon F., Jacob B. et Bimbot F. (2002). « Sirocco : un système ouvert de reconnaissance de la parole ». *Dans Actes des XXIVèmes journées d'étude de la parole (JEP)*, Nancy, France.

Grefenstette G. (1994). « Corpus-Derived First, Second and Third-Order Word Affinities ». *Dans Proc. of EURALEX*, pages 279–290.

Grefenstette G. et Teufel S. (1995). « Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations ». *Dans In Proc. of 7th European Chapter of the Association for Computational Linguistics*, pages 98–103, Dublin, Ireland.

Gretter R. et Riccardi G. (2001). « On-line Learning of Language Models with Word Error Probability Distributions ». *Dans Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 557–560, Salt Lake City, Utah.

Guinaudeau C., Gravier G. et Sébillot P. (2010). « Improving ASR-based Topic Segmentation of TV Programs with Confidence Measures and Semantic Relations ». *Dans*

11th Annual Conf. of the Intl Speech Communication Association, Interspeech'10, Makuhari, Japon.

Gutkin A. (2000). « Log-Linear Interpolation of Language Models ». Mémoire de D.E.A., University of Cambridge.

Ha L. Q., Sicilia-Garcia E. I., Ming J. et Smith F. J. (2002). « Extension of Zipf's Law to Words and Phrases ». *Dans Proc. of the 19th Intl Conf. on Computational Linguistics (COLING)*, volume 1, page 315 ?320, Taipei, Taiwan.

Hart P., Nilsson N. et Raphael B. (1968). « A Formal Basis for the Heuristic Determination of Minimum Cost Paths ». *IEEE Trans. on Systems Science and Cybernetics (SSC4)*, 2:100–107.

Haton J., Cerisara C., Fohr D., Laprie Y. et Smaïli K. (2006). *Reconnaissance Automatique de la Parole Du signal à son interprétation.*

Hearst M. A. (1997). « TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages ». *Computational Linguistics*, 23(1):33–64.

Hetherington I. L. (1995). *A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding.* Thèse de doctorat, MIT, Cambridge, MA, USA.

Huet S. (2007). *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole.* Thèse de doctorat, University of Rennes 1.

Huet S., Gravier G. et Sébillot P. (2007). « Morphosyntactic Processing of N-best Lists for Improved Recognition and Confidence Measure ». *Dans Proceeding of European Conf. on Speech Communication and Technology (Interspeech)*, pages 1741–1744, Antwert, Belgium.

Huet S., Gravier G. et Sébillot P. (2010). « Morpho-Syntactic Post-Processing of N-best Lists for Improved French Automatic Speech Recognition ». *Computer Speech & Language*, 24(4):663–684.

Ibekwe-SanJuan F. (2001). « Extraction terminologique avec INTEX ». *Dans Proc. of the 4th Annual INTEX Workshop*, Bordeaux, France.

Ibekwe-SanJuan F. et SanJuan E. (2003). « From Term Variants to Research Topics ». *Journal of Knowledge Organization (ISKO), special issue on Human Language Technology.*

Ito A., Kajiura Y., Makino S. et Suzuki M. (2008). « An Unsupervised Language Model Adaptation based on Keyword Clustering and Query Availability Estimation ». *Dans Proc. of Intl Conf. on Audio, Language and Image Processing (ICALIP)*, pages 1412–1418.

Jacquemin C. (1997). *Variation terminologique : reconaissance et acquisition automatique de termes et de leur variantes en corpus.* Thèse de doctorat, Université de Nantes.

James F. (2000). « Modified Kneser-Ney Smoothing of n-gram Models ». Rapport technique.

Jelinek F. (1976). « Continuous Speech Recognition by Statistical Methods ». *Proc. of the IEEE*, 64(4):532–556.

Jelinek F. (1998). *Statistical Methods for Speech Recognition*. The MIT Press.

Ji G. et Bilmes J. (2005). « Dialog Act Tagging Using Graphical Models ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP).*, volume 1, pages 33 – 36.

Justeson J. et Katz S. (1995). « Technical Terminology : some Linguistic Properties and an Algorithm for Identification in Text ». *Natural Language Engineering*, pages 9–27.

Justo R. et Torres M. I. (2007). *Computer Recognition Systems II*, chapitre Different Approaches to Class-Based Language Models Using Word Segments, pages 421–428.

Katz S. M. (1987). « Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer ». *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

Kaufmann T., Ewender T. et Pfister B. (2009). « Improving Broadcast News Transcription with a Precision Grammar and Discriminative Reranking ». *Dans Proc. of the Intl Conf. on Speech and Language Technology (Interspeech)*, pages 356–359, Brighton,UK.

Kemp T. et Waibel A. (1998). « Reducing the OOV Rate in Broadcast News Speech Recognition ». *Dans Proc. of the 5th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 1839–1842, Sydney, Australia.

Ketabdar H., Hannemann M. et Hermansky H. (2007). « Detection of Out-of-VocabularyWords in Posterior Based ASR ». *Dans Proc. the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 1757–1760, Antwerp, Belgium.

Kilgarriff A. (2001). « Comparing Corpora ». *Intl Journal of Corpus Linguistics*, 6:1–37.

Kilgarriff A. et Grefenstette G. (2003). « Introduction to the Special Issue on the Web as Corpus ». *Computational linguistics*.

Kirchhoff K., Bilmes J., Duh K., Kirchhoff K., Bilmes J. et Duh K. (2008). « Factored Language Models Tutorial ». Rapport technique, University of Washington.

Kirchhoff K. et Yang M. (2005). « Improved Language Modeling for Statistical Machine Translation ». *Dans Proc. of the ACL Workshop on Building and Using Parallel Texts (ParaTex)*, pages 125–128, Morristown, NJ, USA. Association for Computational Linguistics.

Klakow D. (2000). « Selecting Articles from the Language Model Training Corpus ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1695–1698, Istanbul, Turkey.

Klakow D. (2006). « Language Model Adaptation for Tiny Adaptation Corpora ». *Dans Proc. of Intl Conf. on Speech and Language Technology (Interspeech)*.

Kneser R. et Ney H. (1993). « Improved Clustering Techniques for Class-Based Statistical Language Modelling ». *Dans Proc. of the 3rd European Conf. on Speech Communication and Technology (Eurospeech)*, volume 2, pages 973–976, Berlin, Germany.

Kneser R. et Ney H. (1995). « Improved Backing-Off for M-Gram Language Modeling ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184, Detroit, Michigan, USA.

Kneser R. et Peters J. (1997). « Semantic Clustering for Adaptive Language Modeling ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, page 779, Washington, DC, USA. IEEE Computer Society.

Kneser R., Peters J. et Klakow D. (1997). « Language Model Adaptation using Dynamic Marginals ». *Dans Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1971–1974, Rhodes, Greece.

Kneser R. et Steinbiss V. (1993). « On the Dynamic Adaptation of Stochastic Language Models ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 586–589, Minneapolis, MN, USA.

Kuhn R. et De Mori R. (1990). « A Cache-Based Natural Language Model for Speech Recognition ». *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

Lane I. R., Kawahara T., Matsui T. et Nakamura S. (2005). « Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching ». *IEICE Trans. on Information and Systems*, E88-D(3):446–454.

Langlois D., Brun A., Smaïli K. et Haton J.-P. (2003). « Événements impossibles en modélisation stochastique du langage ». *Traitement Automatique des Langues (TAL)*, 44(1):33–61.

Langlois D., Smaïli K. et Haton J.-P. (2000). « Dealing with Distant Relationships in Natural Language Modelling for Automatic Speech Recognition ». *Dans Proc. of the World MultiConf. on Systemics, Cybernetics and Informatics (SCI)*, volume 6, Orlando, Florida, USA.

Lau R., Rosenfeld R. et Roukos S. (1993). « Trigger-based Language Models : a Maximum Entropy Approach ». *Dans Proc. of IEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 45–48, Los Alamitos, CA, USA. IEEE Computer Society.

Lecorvé G. (2007). « Adaptation thématique d'un système de transcription automatique de la parole ». Mémoire de D.E.A., IRISA / INSA de Rennes.

Lecorvé G., Gravier G. et Sébillot P. (2008a). « On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems ». *Dans* (ELRA) E. L. R. A., éditeur : *Proc. of the 6th Intl Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Lecorvé G., Gravier G. et Sébillot P. (2008b). « An Unsupervised Web-based Topic Language Model Adaptation Method ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech,and Signal Processing (ICASSP)*, pages 5081–5084, Las Vegas, NV, USA.

Lecorvé G., Gravier G. et Sébillot P. (2009). « Constraint Selection for Topic-based MDI Adaptation of Language Models ». *Dans Proc. of the Intl Conf. on Speech and Language Technology (Interspeech)*, pages 368–371, Brighton, UK.

Lecouteux B., Linarès G., Estève Y. et Gravier G. (2008). « Generalized Driven Decoding for Speech Recognition System Combination ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1549–1552.

Lefevre F., Gauvain J.-L. et Lamel L. (2001). « Towards Task-Independent Speech Recognition ». *Dans Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 521–524 vol.1.

Levenshtein V. I. (1966). « Binary Codes Capable of Correcting Deletions, Insertions, and Reversals ». *Soviet Physics Doklady*, 10(8):707–710.

Li X., Singh R. et Stern R. M. (2002). « Combining Search Spaces of Heterogeneous Recognizers for Improved Speech Recognition ». *Dans Proc. of the Intl Conf. on Speech and Language Processing (ICSLP)*, pages 405–408, Denver, Colorado, USA.

Mahajan M., Beeferman D. et Huang X. D. (1999). « Improved Topic-Dependent Language Modeling using Information Retrieval Techniques ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1.

Mangu L., Brill E. et Stolcke A. (2000). « Finding Consensus in Speech Recognition : Word Error Minimization and other Applications of Confusion Networks ». *Computer Speech and Language*, 14(4):373–400.

Manning C. D. et Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Marandin J.-M. (1988). « A propos de la notion de thème de discours ». *Langue Française*, 78:67–87.

Marin M. A., Feldman S., Ostendorf M. et Gupta M. (2009). « Filtering Web Text to Match Target Genres ». *Dans IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 0, pages 3705–3708, Los Alamitos, CA, USA. IEEE Computer Society.

Martin S., Liermann J. et Ney H. (1998). « Algorithms for Bigram and Trigram Word Clustering ». *Speech Communication*, 24(1):19–37.

Martins C., Texeira A. et Neto J. (2006). « Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News Transcription System ». *Dans Proc. of the Spoken Language Technology Workshop*, pages 146–149.

Moreau F. et Claveau V. (2006). « Extension de requêtes par relations morphologiques acquises automatiquement ». *Dans Actes de la 3ème Conférence en Recherche d'Informations et Applications (CORIA)*.

Mou X., Seneff S. et Zue V. (2002). « Integration of Supra-Lexical Linguistic Models with Speech Recognition Using Shallow Parsing and Finite State Transducers ». *Dans Proc. of the 7th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 1289–1292, Denver, Colorado, USA.

Nanjo H., University R. et Kawahara T. (2005). « A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding ». *Dans In Proc. of IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1053–1056.

Naturel X., Gravier G. et Gros P. (2006). « Fast Structuring of Large Television Streams Using Program Guides ». *Dans Adaptive Multimedia Retrieval*, pages 222–231.

Ney H. (1991). « Dynamic Programming Parsing for Context-Free Grammars in Continuous Speech Recognition ». *IEEE Trans. on Signal Processing*, 39(2):336–340.

Niesler T. et Willett D. (2002). « Unsupervised Language Model Adaptation for Lecture Speech Transcription ». *Dans in Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1413–1416.

Niesler T. R. et Woodland P. C. (1996). « A Variable-Length Category-Based N-Gram Language Model ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 164–167, Atlanta, Georgia, USA.

Nisimura R., Komatsu K., Kuroda Y., Nagatomo K., Lee A., Saruwatari H. et Shikano K. (2001). « Automatic N-Gram Language Model Creation from Web Resources ». *Dans Proc. of the European Conf. on Speech Communication and Technology (Eurospeech)*, pages 2127–2130.

Ogata J., Goto M. et Eto K. (2007). « Automatic Transcription for a Web 2.0 Service to Search Podcasts ». *Dans in Proc. of the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 2617–2620, Antwerp, Belgium.

Oger S., Linarès G., Bechet F. et Nocera P. (2008). « On-Demand New Word Learning using World Wide Web ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4305–4308.

Oger S., Popescu V. et Linarès G. (2010). « Combination of Probabilistic and Possibilistic Language Models ». *Dans Proc. of the 8th Annual Conf. of the Intl Speech Communication Association (Interspeech)*.

Ohtsuki K., Hiroshima N., Oku M. et Imamura A. (2005). « Unsupervised Vocabulary Expansion for Automatic Transcription of Broadcast News ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1021–1024.

Palmer D. D. et Ostendorf M. (2005). « Improving Out-of-Vocabulary Name Resolution ». *Computer Speech & Language*, 19(1):107–128.

Palmer D. D., Ostendorf M. et Burger J. D. (2000). « Robust Information Extraction from Automatically Generated Speech Transcriptions ». *Speech Communication*, 32(1-2):95–109. Accessing Information in Spoken Audio.

Rabiner L. (1989). « A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition ». *Proc. of the IEEE*, 77(2):257–285.

Rabiner L. et Juang B. (1993). *Fundamentals of Speech Recognition*. Prentice hall Englewood Cliffs, New Jersey.

Ramshaw L. A. et Marcus M. P. (1995). « Text Chunking using Transformation-Based Learning ». *Dans Proc. of the ACL Third Workshop on Very Large Corpora*, pages 82–94.

Rastier F. (1995). *L'analyse thématique des données textuelles. L'exemple des sentiments*, chapitre La sémantique des thèmes - ou le voyage sentimental, pages 223–249. Didier.

Roark B., Saraçlar M. et Collins M. (2007). « Discriminative N-gram Language Modeling ». *Computer Speech & Language*, 21(2):373–392.

Roche M., Heitz T., Matte-Tailliez O. et Kodratoff Y. (2004). « EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés ». *Dans Actes des Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, volume 2, pages 946–956.

ROSENFELD R. (1994). *Adaptive Statistical Language Modeling : A Maximum Entropy Approach*. Thèse de doctorat, School of Computer Science, Carnegie Mellon University.

ROSENFELD R. (1995). « Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data ». *Dans Proc. of the European Conf. on Speech Technology (Eurospeech)*, volume 2, pages 1763–1766.

ROSENFELD R. (1996). « A Maximum Entropy Approach to Adaptive Statistical Language Modeling ». *Computer, Speech and Language*, 10:187–228.

ROSENFELD R. (2000). « Two Decades of Statistical Language Modeling : Where do we Go from Here ? ». *Proc. of the IEEE*, 88(8):1270–1278.

ROSENFELD R., CHEN S. F. et ZHU X. (2001). « Whole-Sentence Exponential Language Models : a Vehicle for Linguistic-Statistical Integration ». *Computer Speech & Language*, 15(1):55–73.

ROSSIGNOL M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat, Université de Rennes 1.

SALTON G. (1989). *Automatic Text Processing : the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc.

SAMUELSSON C. et REICHL W. (1999). « Class-Based Language Model for Large-Vocabulary Speech Recognition Extracted from Part-of-Speech Statistics ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 537–540, Phoenix, Arizona, USA.

SAPORTA G. (1990). *Probabilités, analyses des données et statistique*, chapitre Les tests statistiques, pages 317–358. Éditions Technip.

SCHARENBORG O., SENEFF S. et BOVES L. (2007). « A Two-Pass Approach for Handling Out-of-Vocabulary Words in a Large Vocabulary Recognition Task ». *Computer Speech & Language*, 21(1):206–218.

SCHWARM S. E., BULYKO I. et OSTENDORF M. (2004). « Adaptive Language Modeling with Varied Sources to Cover New Vocabulary Items ». *IEEE Trans. on Speech and Audio Processing*, 12(3):334–342.

SCHWENK H. et GAUVAIN J.-L. (2002). « Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition ». *Dans Proc. of IEEE Intl Conf on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 765–768.

SCHWENK H. et GAUVAIN J.-L. (2004). « Neural Network Language Models for Conversational Speech Recognition ». *Dans Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 1215–1218, Jeju Island, South Korea.

SEBASTIANI F. (2002). « Machine Learning in Automated Text Categorization ». *ACM Computing Surveys (CSUR)*, 34(1):1–47.

SEIDE F., YU P., MA C. et CHANG E. (2004). « Vocabulary-independent search in spontaneous speech ». *Dans Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–253–6 vol.1.

SETHY A., GEORGIOU P. G. et NARAYANAN S. (2005). « Building Topic Specific Language Models from Webdata Using Competitive Models ». *Dans Proc. of 9th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1293–1296, Lisboa, Portugal.

Seymore K. et Rosenfeld R. (1997). « Using Story Topics for Language Model Adaptation ». *Dans Proc. of the 5th European Conf. on Speech Communication and Technology (Eurospeech)*, pages 1987–1990, Rhodes, Greece.

Singhal A. K. (1997). *Term Weighting Revisited*. Thèse de doctorat, Cornell University, Ithaca, NY, USA.

Smadja F. (1993). « Retrieving Collocations from Text : Xtract ». *Computational Linguistics*, 19(1):143–177.

Smaili K., Jamoussi S., Langlois D. et Haton J.-P. (2004). « Statistical feature language model ». *Dans Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 1357–1360.

Souvignier B., Kellner A., Rueber B., Schramm H. et Seide F. (2000). « The Thoughtful Elephant : Strategies for Spoken Dialog Systems ». *IEEE Transactions on Speech and Audio Processing*, 8(1):51–62.

Stolcke A. (2002). « SRILM - An Extensible Language Modeling Toolkit ». *Dans Proc. of the 7th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, Colorado, USA.

Strzalkowski T. (1995). « Natural Language Information Retrieval ». *Information Processing & Management*, 31(3):397–417. The Second Text Retrieval Conf. (TREC-2).

Suzuki M., Kajiura Y., Ito A. et Makino S. (2006). « Unsupervised Language Model Adaptation based on Automatic Text Collection from WWW ». *Dans Proc. of the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 2202–2205.

Tam Y.-C. et Schultz T. (2006). « Unsupervised Language Model Adaptation Using Latent Semantic Marginals ». *Dans Proc. of the 7th Annual Conf. of the Intl Speech Communication Association (Interspeech)*, pages 2206–2209, Antwert, Belgium.

Tam Y.-C. et Schultz T. (2007). « Correlated Latent Semantic Model for Unsupervised LM Adaptation ». *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE Intl Conf. on*, 4:IV–41–IV–44.

Tür G. et Stolcke A. (2007). « Unsupervised Language Model Adaptation for Meeting Recognition ». *Dans Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

Vaufreydaz D. (2002). *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Thèse de doctorat, Université Joseph Fourier, Grenoble, France.

Vaufreydaz D., Akbar M. et Rouillard J. (1999). « Internet Documents : a Rich Source for Spoken Language Modeling ». *Dans Proc. of the IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, pages 277–280, Keystone, Colorado, USA.

Vergyri D., Kirchhoff K., Duh K. et Stolcke A. (2004). « Morphology-Based Language Modeling for Arabic Speech Recognition ». *Dans Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP)*, pages 2245–2248, Jeju Island, South Korea.

Viterbi A. J. (1967). « Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm ». *IEEE Trans. on Information Theory*, 13(2):260–269.

Weintraub M., Aksu Y., Dharanipragada S., Khudanpur S., Ney H., Prange J., Stolcke A., Jelinek F. et Shriberg L. (1996). « LM95 Project Report : Fast Training and Portability ». Research notes.

Wessel F., Schlüter R., Macherey K. et Ney H. (2001). « Confidence Measures for Large Vocabulary Continuous Speech Recognition ». *IEEE Trans. on Speech and Audio Processing*, 9(3):288–298.

Witten I. H. et Bell T. C. (1991). « The Zero-Frequency Problem : Estimating the Probabilities of Novel Events in Adaptive Text Compression ». *IEEE Trans. on Information Theory*, 37(4):1085–1094.

Xu P., Chelba C. et Jelinek F. (2002). « A study on richer syntactic dependencies for structured language modeling ». *Dans Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 191–198, Morristown, NJ, USA. Association for Computational Linguistics.

Xu P. et Jelinek F. (2004). « Random Forests in Language Modeling ». *Dans In Proc. of Conf. on Empirical Methods in Natural Language Processing.*

Yamashita Y., Tsunekawa T. et Mizoguchi R. (1998). « Topic Recognition For News Speech Based On Keyword Spotting ». *Dans In Proc. of the 5th Intl Conf. on Spoken Language Processing (ICSLP)*, numéro 23, Sydney, Australia. ISCA.

Yamazaki H., Iwano K., Shinoda K., Furui S. et Yokota H. (2007). « Dynamic Language Model Adaptation using Presentation Slides for Lecture Speech Recognition ». *Dans Proc. of the Conf. of the Intl Speech Communication Association (Interspeech)*, pages 2349–2352, Antwerp, Belgium.

Yu H., Tomokiyo T., Wang Z. et Waibel A. (2000). « New Developments in Automatic Meeting Transcription ». *Dans in Proc. of the Intl Conf. on Speech and Language Processing (ICSLP).*

Zitouni I. (2007). « Backoff Hierarchical Class N-gram Language Models : Effectiveness to Model Unseen Events in Speech Recognition ». *Computer Speech & Language*, 21(1):88–104.

Zitouni I. et Zhou Q. (2008). « Hierarchical Linear Discounting Class N-gram Language Models : A Multilevel Class Hierarchy Approach ». *Dans In Proc. of the Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4917–4920.

Table des figures

Liste des tables

Index

INSA de Rennes
Service des formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse : Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole

Nom Prénom de l'auteur : LECORVE Gwénolé

Membres du jury : Monsieur ESTEVE
Madame MOENS
Madame SEBILLOT
Monsieur ADDA
Monsieur BECHET
Monsieur GRAVIER

Président du jury : Frédéric Bechet

Date de la soutenance : 24/11/2010

Reproduction de la thèse soutenue :

- [x] Thèse pouvant être reproduite en l'état
- [ ] Thèse ne pouvant être reproduite
- [ ] Thèse pouvant être reproduite après corrections suggérées

Le Directeur,

A. JIGOREL

Rennes, le 24/11/2010

Signature du Président du jury

Résumé

Les systèmes actuels de reconnaissance automatique de la parole (RAP) reposent sur un modèle de langue (ML) qui les aide à déterminer les hypothèses de transcription les plus probables. Pour cela, le ML recense des probabilités de courtes séquences de mots, appelées *n*-grammes, fondées sur un vocabulaire fini. Ces ML et vocabulaire sont estimés une fois pour toutes à partir d'un vaste corpus de textes traitant de sujets variés. En conséquence, les systèmes actuels souffrent d'un manque de spécificité lorsqu'il s'agit de transcrire des documents thématiquement marqués.

Pour pallier ce problème, nous proposons un nouveau processus d'adaptation thématique non supervisée du ML et du vocabulaire. Sur la base d'une première transcription automatique d'un document audio, ce processus consiste à récupérer sur Internet des textes du même thème que le document, textes à partir desquels nous réestimons le ML et enrichissons le vocabulaire. Ces composants adaptés servent alors à produire une nouvelle transcription dont la qualité est espérée meilleure. Ce processus est particulièrement original car il se préserve de toute connaissance *a priori* sur les éventuels thèmes des documents à transcrire et il intègre des techniques de traitement automatique des langues.

De plus, nous apportons des contributions pour chaque étape du processus. Tout d'abord, étant donnée la transcription initiale d'un document audio, nous avons aménagé le critère *tf-idf*, issu du domaine de la recherche d'information, aux spécificités de l'oral afin de caractériser le thème du document par des mots-clés extraits automatiquement. *Via* un moteur de recherche sur Internet, ces mots-clés nous permettent de récupérer des pages Web que nous filtrons afin d'assurer leur cohérence thématique avec le document audio. Ensuite, nous avons proposé une nouvelle technique de réestimation thématique du ML. En extrayant des mots et séquences de mots spécifiques au thème considéré à partir des corpora Web, nous utilisons le cadre de l'adaptation par minimum d'information discriminante pour ne modifier que les probabilités des *n*-grammes propres au thème, laissant les autres probabilités inchangées. Enfin, nous montrons également que les corpora extraits du Web peuvent servir à repérer des mots hors vocabulaire spécifiques aux thèmes. Nous proposons une technique originale qui permet d'apprendre ces nouveaux mots au système et, notamment, de les intégrer dans le ML en déterminant automatiquement dans quels *n*-grammes ils sont susceptibles d'apparaître. Pour cela, chaque nouveau mot est assimilé à d'autres, déjà connus du système, avec lesquels il partage une relation paradigmatique. Nos expériences, menées sur des émissions d'actualités radiodiffusées, montrent que l'ensemble de ces traitements aboutit à des améliorations significatives du taux de reconnaissance d'un système de RAP.

Abstract

Current automatic speech recognition (ASR) systems are based on language models (LM) which gather word sequence probabilities (*n*-gram probabilities) and assist the system in discriminating utterances with the highest likelihood. In practice, these *n*-gram probabilities are estimated once and for all on large multi-topic corpora based on a fixed, though large, general-purpose vocabulary. Hence, current systems suffer from a lack of specificity when dealing with topic-specific spoken documents.

To circumvent this problem, we propose to modify the LM and the vocabulary through a new unsupervised topic-based adaptation scheme. Based on the sole automatic transcription of a thematically consistent broadcast segment, the process consists in automatically retrieving topic-specific texts on the Internet from which the LM probabilities are re-estimated and the vocabulary is enriched. By running a new transcription process, the use of these adapted components is finally expected to improve the segment recognition accuracy. This work is especially original since it avoids using any *a priori* knowledge about encountered topics and it integrates natural language processing techniques.

In addition, we brought contributions to each step of the adaptation process. First, given a first-pass automatic transcript segment, we propose to adapt indexing methods from the information retrieval domain, namely *tf-idf*, to the specifics of automatic transcription (no case, potentially erroneous words, *etc.*) in order to characterize the encountered topic by a set of keywords. By submitting these keywords to Web search engines, Web pages are then retrieved and thematically filtered to guarantee a good topic similarity with the transcript segment. Second, we developed an original topic-based LM re-estimation technique based on the minimum discrimination information LM adaptation framework and on topic-specific words and phrases automatically extracted from Web corpora. This enables us to exclusively adapt LM *n*-gram probabilities related the topic of the segment, while other, general-purpose, n-gram probabilities are kept untouched. Third, topic-specific Web corpora can be used to spot out-of-vocabulary topic-specific words to be added to the ASR system vocabulary and LM. Whereas adding such words into the vocabulary is straightforward, their integration into a pre-existent LM is more complex. We thus proposed to achieve this task by building *n*-grams for each new word thanks to its paradigmatic relations with other words and thanks to the combined information about the usage of these latter words in the pre-existent LM. Experiments done on French-speaking broadcast news show that our whole topic-specific adaptation process yields significant recognition accuracy improvements of an ASR system.

N° d'ordre : D10-28

Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coësmes . CS 14315 . F-35043 Rennes Cedex
Tel : 02 23 23 82 00 – Fax : 02 23 23 83 96