Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole

Stéphane Huet

N° d'ordre: 3653

THÈSE

présentée

devant l'Université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention INFORMATIQUE

par

Stéphane HUET

Équipes d'accueil : TEXMEX & METISS - IRISA
École Doctorale : MATISSE
Composante universitaire : IFSIC

Titre de la thèse :

Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole

Soutenue le 11 décembre 2007 devant la commission d'examen

Président :	Kamel	SMAÏLI	Université de Nancy 2, LORIA
Rapporteurs :	Philippe	LANGLAIS	Université de Montréal, RALI
	François	YVON	Université de Paris Sud 11, LIMSI-CNRS
Examinateur :	Yannick	ESTÈVE	Université du Maine, LIUM
Directeurs :	Guillaume	GRAVIER	CNRS, IRISA
	Pascale	SÉBILLOT	INSA de Rennes, IRISA

Remerciements

Je tiens à exprimer en premier lieu toute ma gratitude envers mes deux directeurs de thèse, Pascale et Guillaume. Beaucoup de ce qui a été fait au cours de mes travaux n'auraient pu l'être sans leur aide. J'ai grandement apprécié leur encadrement, que ce soit au niveau de la rédaction — où leurs conseils et leur rigueur m'ont aidé à améliorer la qualité du manuscrit et m'ont fait gagner un temps précieux durant la période intensive d'écriture — ou au niveau des expérimentations — qu'eussé-je fait sans les hypothèses de transcription préparées par Guillaume ? Pour parfaire la description, je désirerais ajouter que leur bonne humeur quotidienne et leur passion pour leurs domaines de recherche : le TAL pour Pascale et la RAP pour Guillaume, qui au fur et à mesure de l'avancée de ma thèse ont évolué vers l'utilisation conjointe du TAL et de la RAP, ont permis de créer une ambiance très conviviale de travail.

Je souhaite également remercier chaque membre du jury pour l'intérêt qu'ils ont tous manifesté pour ce que j'ai pu réaliser. Je remercie ainsi Kamel Smaïli qui m'a fait l'honneur de présider mon jury, François Yvon et Philippe Langlais pour les remarques très instructives qu'ils m'ont prodiguées en qualité de rapporteurs, ainsi que Yannick Estève pour avoir accepté de juger mes travaux.

L'IRISA est un environnement qui permet de faire de la recherche dans de bonnes conditions. J'ai particulièrement apprécié le matériel mis à ma disposition, à la pointe au niveau des espaces de stockage et de la rapidité de calcul. Mes remerciements donc à ceux qui font de ce laboratoire un cadre aussi privilégié. Les dispositifs matériels ne sont tout de même pas tout, quelqu'informaticien que je puisse être, et je n'oublie pas non plus l'ambiance chaleureuse qui règne dans l'équipe TexMex, qui m'a accueilli pendant ces années. Je remercie tous les membres — anciens ou actuels — avec qui j'ai pu notamment discuter à la cafet'. Je tiens tout particulièrement à remercier les jeunes chercheurs en TAL de TexMex, Fabienne et Vincent, qui m'ont apporté des conseils précieux. Je souhaite également remercier Gwénolé pour le travail très utile qu'il a accompli durant son Master recherche et lui exprime mes encouragements dans la réussite de ses travaux de thèse.

Avoir obtenu mon doctorat est un aboutissement d'un long travail, parfois semé de difficultés. Je remercie tous ceux qui ont eu confiance en moi pour me permettre de la réussir. J'ai également une pensée toute particulière pour deux amis de DEA qui, pour des raisons diverses, ont dû retourner dans leur pays respectif, le Cambodge et la Tunisie.

Je remercie enfin mes parents pour leur soutien.

Table des matières

Table des figures

Introduction

La parole est une riche source d'information pour de nombreux documents audio et vidéo, et savoir l'exploiter se révèle primordial dans nombre d'applications. Ainsi pour ne citer que quelques exemples, que ce soit pour rechercher un événement précis dans un ensemble de données, détecter les locuteurs qui interviennent dans une émission, traduire ce que dit quelqu'un ou résumer automatiquement un document, comprendre les informations véhiculées par la voix humaine est fondamental. Dans certains contextes, le recours à des personnes pour effectuer toutes ces tâches s'avère coûteux, voire même impossible lorsque le volume de données à traiter devient gigantesque. Parmi l'ensemble des outils développés pour exploiter automatiquement la voix, les systèmes de reconnaissance automatique de la parole (RAP), dont l'objectif est de transcrire à l'aide d'une machine un signal sonore en texte, jouent un rôle central. Depuis les premiers systèmes dans les années 50, les performances ont beaucoup progressé puisque les meilleurs d'entre eux parviennent maintenant à transcrire les émissions d'information diffusées à la radio avec à peine un peu plus d'un mot sur dix erroné. Cette amélioration est expliquée par l'utilisation de méthodes statistiques efficaces pour identifier les mots associés aux sons prononcés par un locuteur, par la mise à disposition d'un volume toujours plus important de données pour paramétrer les modèles de reconnaissance et par la progression des capacités des ordinateurs. Toutefois, il est souvent reconnu que les méthodes statistiques, malgré leur succès, rencontrent des limites et qu'il faut recourir à des connaissances supplémentaires pour faire progresser encore la qualité des transcriptions [Rosenfeld, 2000a].

Au vu du matériau manipulé, du langage naturel, les connaissances sur la langue semblent des candidats logiques pour guider le décodage des documents audio contenant de la parole. Ainsi, des informations sur la structure des phrases, la forme des mots ou les sujets abordés au cours d'une discussion sont *a priori* des éléments pertinents pour transcrire un document. Paradoxalement, les informations de cet ordre réellement utilisées par la plupart des systèmes de RAP sont réduites. Elles se limitent bien souvent à l'utilisation d'un dictionnaire établissant les différentes manières de prononcer les mots d'un vocabulaire et à la connaissance des probabilités de successions de mots. Cette constatation trouve sa source dans l'histoire de la RAP, illustrée par la phrase prononcée en 1988 par Frederick Jelinek, alors chercheur dans le groupe de parole

d'IBM : « *Anytime a linguist leaves the group the recognition rate goes up.* »[1]. Pendant une longue période, les linguistes et les adeptes du traitement automatique des langues (TAL) ont en effet délaissé les méthodes empiriques permettant d'estimer la probabilité d'observation d'un phénomène à partir d'une grande collection de documents ou corpus, celles-là même qui étaient les plus efficaces pour choisir la meilleure hypothèse de mots dans la RAP. Les travaux du TAL n'ont ainsi pu apporter qu'un gain effectif très limité au niveau de la performance des systèmes de transcription. Aujourd'hui encore, la technique la plus employée en RAP pour modéliser le langage reste les modèles N-grammes, prédisant le mot suivant à partir des $N-1$ qui viennent d'être reconnus. Cette modélisation, bien que très simple, a été améliorée depuis son introduction en RAP par des méthodes de lissage des probabilités mais le principe reste fondamentalement le même.

L'utilisation de connaissances linguistiques au cours du processus de transcription se heurte de fait à plusieurs difficultés. Premièrement, elle nécessite des méthodes d'analyse suffisamment robustes pour fonctionner sur des hypothèses proposées par le système de RAP. Si l'on considère par exemple les connaissances relevant de la syntaxe, des outils doivent être employés pour déterminer la structure de la succession de mots déjà reconnus, de façon à guider le décodage du signal sonore. Le recours à des techniques du TAL semble naturel ; toutefois, celles-ci sont généralement utilisées sur des documents de langue écrite, dont certaines caractéristiques sont très éloignées de celles des sorties produites par un système de RAP. Les transcriptions automatiques sont en effet très bruitées puisqu'elles ne contiennent aucune segmentation linguistique sous forme de ponctuation mais sont en revanche marquées par la présence de phénomènes d'hésitations ou d'erreurs de transcription qui sont sujets à perturber l'analyse d'une succession de mots. De plus, une fois les informations obtenues, encore faut-il savoir les employer dans le processus de transcription. Les méthodes symboliques sont notamment délicates à intégrer dans l'approche probabiliste adoptée par la plupart des systèmes de RAP actuels.

Dans cette thèse, nous nous intéressons à la contribution des connaissances linguistiques pour améliorer la qualité des transcriptions produites par un système de RAP. Nous nous focalisons pour ce faire sur deux directions : l'apport pour la RAP des parties du discours, appelées aussi POS pour *parts of speech* en anglais, et l'adaptation thématique d'un système de transcription. Les POS sont des catégories grammaticales associées à un mot, voire à un groupe de mots dans le cas de locutions. Elles correspondent à des fonctions syntaxiques comme celles de verbe, nom, préposition, remplies par un mot dans un contexte donné. Elles peuvent en outre être complétées par des informations morphologiques sur le genre, le nombre, le cas, le temps ou le mode. Ce type de connaissances a plusieurs intérêts en RAP. Les systèmes de transcription présentent, comme nous l'avons d'ailleurs constaté nous-même pour le français, une part importante d'erreurs d'accord ; les POS accompagnées d'informations sur le genre et le nombre sont donc pertinentes pour traiter cette classe d'erreurs. De plus, ce sont des connaissances qui sont plus simples à extraire que d'autres de plus de haut niveau,

[1] À chaque fois qu'un linguiste quitte le groupe, le taux de reconnaissance augmente.

comme celles conduisant à la production d'un arbre syntaxique. Enfin, la désambiguïsation d'un mot selon sa catégorie grammaticale est une étape préalable à de nombreux traitements. Ainsi, dans le cadre de la segmentation d'un document en fonction des thèmes abordés, une sélection des adjectifs, noms et verbes, *i.e.*, des classes porteuses de sens, peut être nécessaire.

Les premières études visant à prendre en compte les POS dans un système de RAP datent des années 80 [Jelinek, 1990]. Il existe toutefois peu de résultats concluants quant à l'apport de ces étiquettes catégorielles pour améliorer la qualité de la transcription et leur utilisation reste très limitée dans les systèmes de transcription actuels [Rosenfeld, 2000a]. L'originalité de nos travaux repose sur trois points. Nous utilisons tout d'abord les POS de manière à introduire explicitement des connaissances morphologiques en RAP. Nos jeux d'étiquettes sont pour ce faire construits en associant les POS à des informations de genre et de nombre. En outre, nous désambiguïsons les mots à reconnaître en fonction de leur classe grammaticale avant d'utiliser cette nouvelle connaissance au cours du processus de décodage, contrairement à la majorité des études envisageant plutôt l'ensemble des POS possibles associées à un mot donné [Jelinek, 1990 ; Maltese et Mancini, 1992]. Nous montrons que cette désambiguïsation peut être réalisée de manière fiable à partir de l'étiqueteur statistique que nous avons développé [Huet *et al.*, 2006a]. Enfin, nous proposons un nouveau mode de prise en compte des POS par un système de RAP en les intégrant au niveau des groupes de souffle, plutôt que des mots. Nos expériences attestent que cette méthode permet une amélioration significative que ce soit au niveau de la qualité de la transcription produite ou de la détection des erreurs de reconnaissance du système de RAP. Cette amélioration s'est révélée robuste à différents paramètres puisque nous l'avons montrée tant pour de l'anglais que pour du français, et y compris pour des extraits *a priori* plus difficiles à traiter comme ceux contenant beaucoup de marques d'hésitation [Huet *et al.*, 2007].

Contrairement à la première partie de notre travail qui présente une intégration finalisée de nos contributions au sein du système de RAP, la seconde demeure avant tout exploratoire, du fait des problèmes posés et de son originalité. Nous nous intéressons à une autre manière d'améliorer la qualité de la transcription, en cherchant à adapter un système de RAP en fonction des thèmes traités par le document étudié. Ces travaux sont justifiés par le fait qu'il existe une variabilité importante des tournures et du vocabulaire en fonction du sujet abordé ; le langage employé par un commentateur sportif est ainsi très différent de celui utilisé pour décrire l'évolution de la Bourse. Or, les systèmes de RAP conventionnels ont des difficultés à prendre en compte ces différences puisque leurs paramètres sont fixés *a priori* sur un corpus généraliste puis ensuite appliqués à l'identique à différents documents. Bien que séduisante, l'adaptation d'un système de reconnaissance soulève plusieurs problèmes. Elle nécessite tout d'abord d'identifier dans une transcription les parties qui doivent être traitées indépendamment lors de l'adaptation, ce qui suppose de savoir localiser les passages où un document passe d'un sujet à un autre et à quel point un thème peut être considéré comme similaire à un autre. Elle requiert également d'être capable de reconnaître de quoi parlent les locuteurs afin de choisir des corpus pertinents et de réussir à utiliser ces données pour qu'une amélioration du nombre de mots bien reconnus soit constatée. Elle se heurte

enfin aux difficultés rencontrées dans les travaux existant sur l'adaptation des modèles d'un système de RAP pour augmenter en pratique la qualité des transcriptions [Gildea et Hofmann, 1999 ; Rosenfeld, 2000b].

Notre approche consiste à adapter le système de RAP en trois étapes [Huet *et al.*, 2008]. Une première opération consiste à segmenter une transcription produite avec un modèle général en segments homogènes sur le plan thématique. Si la détection des points de changement de thème a souvent été étudiée sur des documents écrits dans le domaine du TAL, elle l'a plus rarement été sur des transcriptions automatiques. Nous contribuons à ce domaine en l'envisageant pour adapter de manière plus fine un système de RAP pour des émissions de plusieurs dizaines de minutes. Nous proposons en outre une méthode novatrice s'appuyant sur différents types d'indices tels que la présence de certaines séquences de mots annonciatrices d'un changement de sujet, la cohésion lexicale ou encore l'apparition de sonals[2]. Dans une deuxième étape, nous construisons des corpus de textes pertinents pour chaque section thématique et qui vont permettre par la suite de modifier le système de transcription. Pour ce faire, nous extrayons des mots-clés jugés caractéristiques du document étudié, ce qui nous permet de ne pas avoir à caractériser *a priori* les thèmes susceptibles d'être abordés par le document à transcrire, contrairement à la plupart des techniques d'adaptation actuelles. Nous employons alors ces termes pour construire des requêtes à un moteur de recherche sur Internet, et obtenir ainsi des pages Web pertinentes vis-à-vis du sujet traité. Les corpus ainsi constitués sont délicats à construire du fait de la grande hétérogénéité des documents disponibles sur Internet ; cette manière originale de procéder nous permet toutefois de disposer d'un ensemble volumineux de données quel que soit le sujet envisagé. Dans une dernière étape, nous modifions le système de reconnaissance pour chaque segment thématique à partir des corpus précédemment formés. Même si notre prototype ne permet pas d'employer de manière séquentielle et totalement automatisée les trois maillons de cette chaîne, de premiers résultats indiquent que la transcription est améliorée lorsque les sections traitées sont suffisamment homogènes [Huet *et al.*, 2008]. Nous analysons en outre les raisons de la faiblesse actuelle de l'approche qui laissent envisager un gain significatif au niveau de la qualité de la transcription.

Organisation du manuscrit

Ce mémoire est articulé autour de trois parties.

Partie I - Contexte d'étude et état de l'art. Le premier chapitre de ce mémoire présente le fonctionnement général d'un système de RAP. Nous y abordons les limitations des méthodes statistiques ordinairement utilisées et les difficultés inhérentes à la transcription de documents sonores. Le deuxième chapitre propose un état de l'art sur l'introduction de connaissances linguistiques en vue d'améliorer la qualité de la transcription. Il indique les types d'informations linguistiques qui ont été envisagés en reconnaissance de la parole et où cette utilisation peut intervenir dans le processus

[2] Équivalent français de *jingles*.

de transcription. Nous exposons également les difficultés d'acquisition automatique de chacune des connaissances à partir de documents oraux produits par un système de RAP. Le troisième chapitre mentionne le cadre expérimental employé pour tester nos différentes contributions, en fournissant des informations sur le corpus étudié et le système de transcription développé à l'IRISA.

Partie II - Informations morpho-syntaxiques et reconnaissance de la parole. Cette seconde partie présente nos contributions sur l'utilisation des POS en RAP. Nous évoquons dans un premier temps (chapitre 4) la désambiguïsation d'une transcription selon les POS, en décrivant l'étiqueteur que nous avons développé. Nous proposons ensuite (chapitre 5) plusieurs méthodes de prise en compte de ces classes catégorielles en montrant expérimentalement qu'elles conduisent à une amélioration de la qualité de la transcription et de la détection des mots mal reconnus.

Partie III - Adaptation thématique de la reconnaissance de la parole. Nous relatons dans cette partie nos apports pour modifier un système de RAP en fonction des thèmes abordés dans une émission. Nous y présentons tout d'abord (chapitre 6) la succession des étapes que avons identifiées comme nécessaires pour adapter le processus de transcription, avant de décrire (chapitre 7) plus précisément les méthodes que nous avons développées pour segmenter thématiquement un document sonore contenant de la parole, en soulignant les difficultés posées par cette opération. Le dernier chapitre expose nos choix pour construire à partir du Web des corpus spécifiques à chaque section thématique, en indiquant les résultats obtenus avec une méthode d'adaptation relativement simple, dans cette partie volontairement plus exploratoire que celle consacrée aux informations morpho-syntaxiques.

Nous concluons finalement nos travaux de recherche, en en proposant une synthèse et en présentant des perspectives de recherche qu'ils offrent.

Première partie

Contexte d'étude et état de l'art

Chapitre 1

Principes de la reconnaissance automatique de la parole

L'objectif de la reconnaissance automatique de la parole (RAP) est d'identifier les mots prononcés dans un signal sonore contenant de la parole. Pour y parvenir, les systèmes de transcription ont recours à divers modèles statistiques pour faire la correspondance entre les caractéristiques acoustiques du signal et les hypothèses de mots les plus probables. Parmi ces modèles, deux sont presque systématiquement utilisés : les modèles de Markov cachés pour modéliser la probabilité qu'un son donné corresponde à un mot du langage et les modèles N-grammes pour évaluer la probabilité de successions de mots. Ces modèles permettent aujourd'hui d'avoir de très bonnes performances pour un nombre croissant d'applications, notamment — pour ne mentionner que deux exemples — pour la dictée vocale et la commande vocale à vocabulaire restreint.

Toutefois, la RAP demeure encore de nos jours un problème difficile, particulièrement lorsque l'on souhaite traiter des types variés de documents, où les diversités de conditions d'enregistrement, d'accent des locuteurs ou bien encore de sujets abordés, peuvent grandement complexifier le processus de transcription. En outre, malgré leur succès, les méthodes statistiques sont basées sur des hypothèses simplificatrices qui les conduisent à produire des transcriptions dans lesquelles subsistent des erreurs.

Ce chapitre est dédié aux principes de fonctionnement de la RAP, en présentant certaines limitations des systèmes actuels pour modéliser certains phénomènes et les principales difficultés auxquelles sont confrontées les méthodes appliquées à la transcription de la parole. Nous décrivons tout d'abord les étapes successives qui permettent à des systèmes de transcrire automatiquement un signal sonore. Nous examinons plus en détail par la suite le modèle de langage, qui a pour vocation d'introduire des connaissances sur la langue au cours du processus de décodage. Après avoir évoqué les méthodes d'évaluation des systèmes de RAP, nous indiquons enfin les difficultés de la transcription liées à la grande variabilité des documents à traiter.

1.1 Modélisation statistique de la reconnaissance automatique de la parole

Dans cette section nous présentons comment les systèmes de RAP basés sur une modélisation statistique parviennent à transcrire la parole. La plupart des décodeurs actuels, dont celui que nous utilisons au cours de nos expérimentations, sont généralement conçus suivant le même principe que celui que nous exposons.

Les valeurs d'entrée d'un système de RAP correspondent à un signal sonore échantillonné ; elles sont analysées de manière à extraire une séquence d'observations acoustiques A. Dans le cadre d'une modélisation statistique du décodage de la parole, la recherche de la succession W des mots prononcés est basée sur un critère MAP (maximum *a posteriori*) :

$$\hat{W} = \arg\max_{W \in \mathcal{L}} P(W|A) \ . \tag{1.1}$$

L'objectif de la transcription de la parole revient alors à rechercher la chaîne $\hat{W}$ la plus probable parmi toutes les séquences possibles d'un langage $\mathcal{L}$, à partir des indices acoustiques A extraits du signal. À l'aide de la formule de Bayes, l'équation (1.1) peut être réécrite sous la forme suivante [Jelinek, 1998] :

$$\hat{W} = \arg\max_{W \in \mathcal{L}} \frac{P(W)P(A|W)}{P(A)} \tag{1.2}$$

où $P(W)$ est la probabilité que W soit prononcée, $P(A|W)$ est la probabilité que le locuteur émette A en souhaitant prononcer les mots W et $P(A)$ est la probabilité que A soit produite. $\hat{W}$ étant estimée en fixant A, $P(A)$ n'intervient pas et (1.2) devient :

$$\hat{W} = \arg\max_{W \in \mathcal{L}} P(W)P(A|W) \ . \tag{1.3}$$

Cette équation met en évidence les différents composants d'un système de RAP (Fig. 1.1) :

- un module extrayant les caractéristiques acoustiques A à partir du signal,
- un dictionnaire de prononciation fixant les mots du langage $\mathcal{L}$ reconnaissables par le système de RAP,
- un *modèle acoustique* (MA) calculant la probabilité d'observation $P(A|W)$ pour des hypothèses W qui sont vraisemblablement associées à A,
- un *modèle de langage* (ML) calculant la probabilité *a posteriori* $P(W)$ pour les hypothèses W choisies par le MA, en fonction de connaissances sur $\mathcal{L}$,
- un décodeur mettant à jour la liste des hypothèses les plus probables W à partir des informations fournies par le dictionnaire de prononciation, le MA et le ML, et ce en parcourant progressivement le signal du début à la fin.

Les sections suivantes décrivent succinctement le fonctionnement de chacun des composants ; pour une description plus complète, on pourra se référer par exemple à [Jurafsky et Martin, 2008]. Elles exposent en outre les types de données que peuvent produire les systèmes de RAP en sortie et indiquent enfin l'intérêt des mesures de confiance pour la transcription.

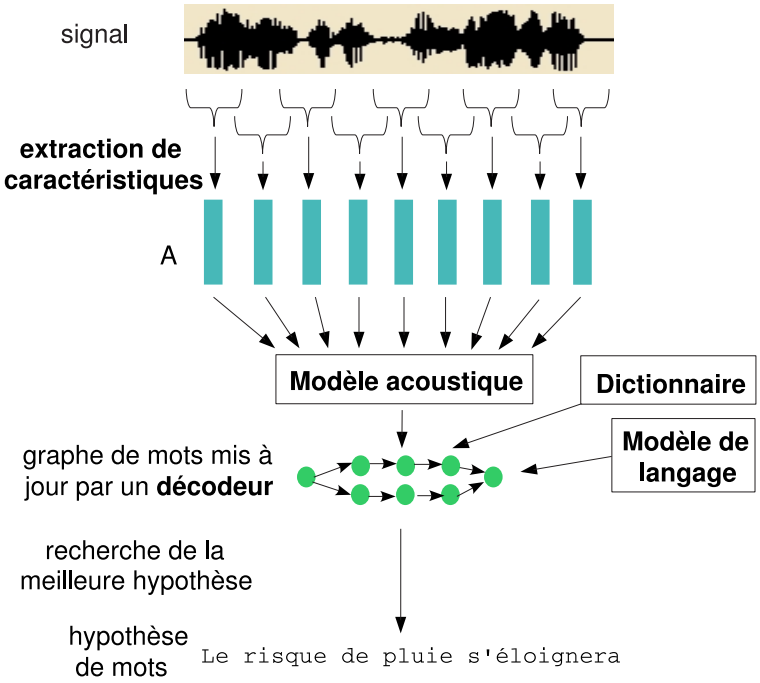

FIG. 1.1: *Constituants d'un système de transcription*

1.1.1 Extraction de caractéristiques

Le signal sonore à analyser étant susceptible de contenir, en sus de la parole, des zones de silence, de bruit ou encore de musique, un premier traitement consiste à partitionner le flux audio pour ne retenir que les portions contenant de la parole. Le signal est ensuite segmenté en ce que l'on appelle des *groupes de souffle*, en utilisant comme délimiteurs des pauses silencieuses suffisamment longues, *i.e.*, de l'ordre de 0,3 seconde. L'intérêt de cette segmentation est de se restreindre à l'analyse d'un signal de taille raisonnable par rapport aux capacités de calculs des modèles du système de RAP ; dans la suite du processus de transcription, l'analyse se fait séparément pour chaque groupe de souffle.

Une fois le flux audio ainsi découpé, entre en jeu la phase d'extraction de caractéristiques proprement dite. Son objectif est d'obtenir une succession de valeurs numériques suffisamment informatives sur le plan acoustique pour pouvoir reconnaître des mots par la suite. Pour ce faire, le signal, présentant généralement une fluctuation rapide au cours du temps, est découpé en fenêtres d'études de 10, 15 ou 20 millisecondes. Pour chaque fenêtre d'analyse i sont ensuite calculés des coefficients a_i représentant l'enveloppe spectrale du signal. Le but consiste ici à extraire les propriétés discriminant les sons caractéristiques de la voix dans les bandes de fréquences utilisées par le système auditif humain. À l'issue de cette étape, le flux audio devient une succession de vecteurs $A = a_1 \dots a_t$ dits acoustiques.

adorateurs	a d O R a t 9 R z
adorateurs	a d O R a t 9 R
adoration	a d O R a s j o~
adore	a d O R @
adore	a d O R

FIG. 1.2: *Extrait d'un dictionnaire de prononciation*

1.1.2 Dictionnaire de prononciation

Lors de la conception d'un système de RAP, il est nécessaire de fixer un vocabulaire $\mathcal{V}$ représentant l'ensemble des mots pouvant être reconnus, ce qui restreint le langage $\mathcal{L}$ des séquences reconnaissables à $\mathcal{V}^*$. Une caractéristique importante de la RAP est ainsi qu'elle se base sur un nombre fini de mots, généralement déterminés en ne retenant que les plus fréquents d'un corpus représentatif du type de documents que l'on souhaite transcrire.

D'un point de vue plus linguistique, chaque mot de $\mathcal{V}$ désigne en réalité un *mot-forme* défini par son orthographe, le terme mot étant porteur d'ambiguïté [Polguère, 2003]. Ainsi, deux flexions ou dérivations d'un même lemme[1] (*e.g.* « *mange* » et « *manges* ») sont considérées comme deux mots différents. En revanche, deux homographes[2] appartenant à deux catégories différentes (*e.g.* « *mérite* [VERBE] » et « *mérite* [NOM] ») ou deux sens différents (*e.g.* « *avocat* [AUXILIAIRE DE JUSTICE] » et « *avocat* [FRUIT COMESTIBLE] ») ne représentent pas la même lexie[3] mais possèdent pourtant le même mot-forme. Dans la suite de ce mémoire, nous nous tiendrons au vocabulaire du domaine de la RAP en employant le terme *mot*, en gardant à l'esprit qu'il désigne en réalité un *mot-forme*.

Outre la liste des constituants de $\mathcal{V}$, le lexique fixe également les prononciations autorisées pour chaque mot. Comme chaque élément du vocabulaire est susceptible d'être prononcé de différentes manières, selon son prédécesseur et son successeur dans le groupe de souffle, ou tout simplement selon les habitudes du locuteur, il peut y avoir plusieurs entrées dans le dictionnaire pour un même mot. Les indications sont données au moyen des *phonèmes*[4] caractéristiques de la prononciation. Sur la figure 1.2 montrant un extrait d'un dictionnaire, les phonèmes sont transcrits dans le système de représentation SAMPA.

Dans la suite de ce mémoire, la taille du vocabulaire d'un système de RAP sera notée $|\mathcal{V}|$. Elle correspond au nombre d'entrées lexicales dans le dictionnaire de prononciations,

[1] Forme canonique d'un mot-forme dépendant de sa catégorie et obtenu grâce à une analyse morphologique. Le lemme d'un verbe conjugué sera par exemple son infinitif, celui d'un adjectif, sa forme masculin singulier.

[2] Mots de catégorie ou de sens différents, correspondant au même mot-forme.

[3] Unité du lexique correspondant à un sens particulier et associé à un mot-forme ou à une expression.

[4] Unité linguistique associée à un type de prononciation d'une langue donnée. Le phonème final /p/ pourra par exemple être prononcé en français par le phone [b] dans l'expression « *grippe aviaire* » et par le phone [p] dans « *grippe du poulet* ».

deux variantes de prononciation d'un même mot étant comptabilisées comme une seule entité.

1.1.3 Modèle acoustique

Le rôle du MA consiste à associer aux vecteurs acoustiques, qui sont des vecteurs numériques, un ensemble d'hypothèses de mots, autrement dit des symboles. L'outil le plus utilisé pour la modélisation du MA est le modèle de Markov caché (MMC), celui-ci ayant en effet montré dans la pratique son efficacité pour reconnaître la parole. Même s'il présente quelques limitations pour prendre en compte certaines caractéristiques du signal, comme la durée ou la dépendance des observations acoustiques successives, ce type de modèle offre un cadre mathématique bien défini pour déterminer les probabilités $P(A|W)$ [Rabiner, 1989]. Les MMC utilisés au sein du MA sont d'ordre 1 ; ils considèrent que l'état présent q_t dépend uniquement de l'état q_{t-1} où l'on était précédemment. Ce sont en outre des MMC gauche-droite (appelés aussi des MMC de Bakis) puisqu'ils interdisent les transitions vers un état antérieur. Le calcul de $P(A|W)$ est effectué à partir de deux types de paramètres :

- les probabilités de transition $P(q_t = S_j|q_{t-1} = S_i)$ d'un état S_i vers un état S_j, où chaque état représente un symbole ;
- les probabilités d'émission $P(o_t|q_t = S_j)$ de l'observation o_t dans l'état S_j à l'instant t, où les observations sont des vecteurs acoustiques.

Les MA font intervenir trois niveaux de MMC (Fig. 1.3). Ils cherchent dans un premier temps à reconnaître les types de son, autrement dit à identifier les *phones*[5] parmi ceux possibles dans le langage $\mathcal{L}$. Pour ce faire, ils modélisent généralement chaque phone par un MMC à trois états représentant ses début, milieu et fin. Les états q_t sont alors associés à un *sous-phone*[6], tandis que les probabilités d'émission sont souvent modélisées par des gaussiennes. Le deuxième niveau de MMC représente des mots ; c'est ici qu'entre en jeu le dictionnaire de prononciation. Ce dernier est utilisé dans le MA en représentant l'ensemble des entrées sous la forme d'un même arbre lexical. Comme nous l'avons vu à la section 1.1.2, le dictionnaire indique des phonèmes et non des phones. Pour faire la liaison entre les deux premiers niveaux du MA, les phonèmes sont convertis en phones au moyen de règles de transformation dépendant de leur contexte d'apparition. Le troisième niveau modélise enfin la succession des mots W au sein d'un groupe de souffle et peut alors incorporer les connaissances apportées par le ML sur W. Pour établir ce MMC équivalent à un graphe de mots, le MMC correspondant à l'arbre lexical est dupliqué à chaque fois que le MA effectue l'hypothèse qu'un nouveau mot a été reconnu.

L'apprentissage des paramètres du MA est réalisé par maximum de vraisemblance, en utilisant l'algorithme de Baum-Welch [Baum *et al.*, 1970] ou une approximation, l'algorithme forcé de Viterbi [Jurafsky et Martin, 2008]. Les corpus utilisés pour ce faire comportent un flux audio, accompagné de sa transcription sous forme de mots.

[5] Sons prononcés par un locuteur et définis par des caractéristiques acoustiques précises. Un phone peut ainsi être vu comme la réalisation physique d'un phonème, qui est lui une abstraction cognitive.

[6] Un des trois extraits successifs d'un phone.

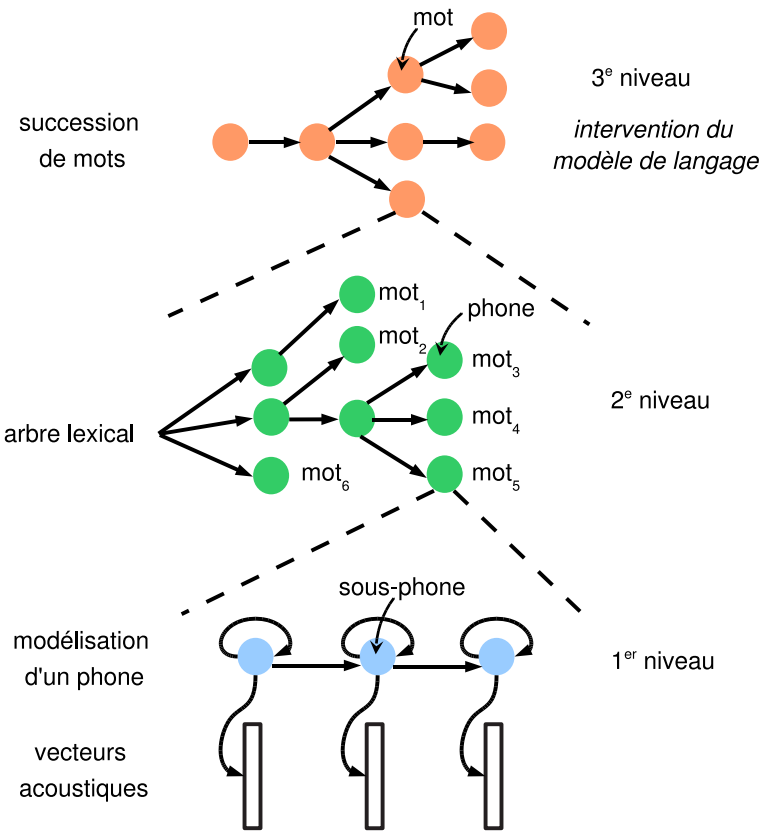

FIG. 1.3: *Niveaux de modélisation du modèle acoustique*

L'alignement du texte avec le signal est réalisé automatiquement durant le processus d'apprentissage, sans qu'il soit nécessaire d'effectuer une transcription phonétique en précisant les limites temporelles de chaque phone.

1.1.4 Modèle de langage

Le ML a pour objectif de guider la recherche des séquences de mots les plus probables. Il détermine à cette fin les probabilités $P(W)$ de différentes hypothèses, en posant $W = w_1^n = w_1 \dots w_n$, où w_i est le mot de rang i de la séquence W. Le calcul est généralement décomposé de la manière suivante :

$$P(w_1^n) = P(w_1) \prod_{i=2}^{n} P(w_i | w_1 \dots w_{i-1}) \tag{1.4}$$

où $P(w_1)$ représente la probabilité d'observer le mot w_1 et $P(w_i|w_1 \dots w_{i-1})$ celle de rencontrer le mot w_i après la séquence $w_1 \dots w_{i-1}$. Ces deux probabilités sont obtenues

respectivement à l'aide des égalités :

$$P(w_1) = \frac{C(w_1)}{\sum_{w \in \mathcal{V}} C(w)} \tag{1.5}$$

$$P(w_i|w_1^{i-1}) = \frac{C(w_1^i)}{\sum_{w_i \in \mathcal{V}} C(w_1^i)} = \frac{C(w_1^i)}{C(w_1^{i-1})} \tag{1.6}$$

où $C(w_1)$ et $C(w_1^i)$ représentent les nombres d'occurrences respectifs du mot w_1 et de la séquence de mots w_1^i dans un corpus d'apprentissage textuel, de taille souvent beaucoup plus grande que celui qui a été utilisé pour l'évaluation des paramètres du MA.

Malheureusement, pour prédire la suite de mots w_1^n, le nombre des paramètres $P(w_1)$ et $P(w_i|w_1^{i-1})$ à estimer pour le ML augmente de manière exponentielle avec n. Dans le but de réduire ce nombre, $P(w_i|w_1^{i-1})$ est souvent modélisé par un *modèle N-grammes*, *i.e.*, une chaîne de Markov d'ordre $N-1$ (avec $N > 0$), à l'aide de l'équation suivante :

$$P(w_i|w_1^{i-1}) \approx P(w_i|w_{i-N+1}^{i-1}) \ . \tag{1.7}$$

Cette équation indique que chaque mot w_i peut être prédit à partir des $N-1$ mots précédents. Pour $N = 2, 3$ ou 4, on parle respectivement de modèle *bigrammes*, *trigrammes* ou *quadrigrammes*. Pour $N = 1$, le modèle est dit *unigramme* et revient à estimer $P(w_i)$. Généralement, ce sont les modèles bigrammes, trigrammes et quadrigrammes qui sont utilisés dans les ML des systèmes de RAP.

Certains mots apparaissent davantage en début de groupe de souffle que d'autres. Pour tenir compte de cette caractéristique, « <s> » est ajouté en pratique devant w_1, afin de priviliégier les séquences « <s> w_1 ... » les plus probables. De la même manière, les fins de groupe de souffle ayant tendance à apparaître après certaines séquences de mots plutôt que d'autres, un nouveau terme « </s> » est positionné après w_n. Une autre justification de l'ajout de « </s> », proposée dans [Chen et Goodman, 1998], est que sans lui, la probabilité de toutes les séquences possibles $\sum_{W \in \mathcal{L}} P(W)$ ne serait pas égale à 1 mais à l'infini, tandis que la somme des probabilités de toutes les chaînes d'une longueur donnée $\sum_{W \in \mathcal{L}, |W|=n} P(W)$ serait, elle, égale à 1. De manière à accroître la précision des calculs, « <s> » et « </s> » sont donc généralement intégrés aux groupes de souffle. La probabilité de la séquence « *il pleut* » par un ML bigrammes devient par exemple :

$$P(\text{« } il\ pleut \text{ »}) = P(\text{« } il \text{ »}|\text{« <s> »})\, P(\text{« } pleut \text{ »}|\text{« } il \text{ »})\, P(\text{« </s> »}|\text{« } pleut \text{ »}) \ . \tag{1.8}$$

1.1.5 Décodeur

Le décodeur cherche à construire un graphe des mots ayant pu être prononcés dans le signal à analyser, à partir des informations transmises par le dictionnaire de prononciation, le MA et le ML. La principale difficulté rencontrée à ce niveau est que l'espace de recherche devient rapidement immense dès que les groupes de souffle à transcrire contiennent plus d'un mot. Pour ne pas avoir à examiner toutes les séquences possibles des mots de $\mathcal{L}$, avec toutes leurs variantes de prononciation, le graphe de mots est

construit au fur et à mesure de la progression dans l'analyse du signal, en ne retenant que les hypothèses les plus probables ; ce graphe correspond au troisième niveau du MA (*cf.* section 1.1.3). Afin de limiter les risques d'écarter trop tôt une proposition correcte, le décodeur prend en compte au plus tôt les probabilités calculées par le ML. En pratique, cela consiste à dupliquer l'arbre lexical lorsque l'on suppose une fin de mot et à ajouter un arc entre deux mots consécutifs portant la probabilité $P(w_i|w_1 \dots w_{i-1})$ du ML [Ortmanns *et al.*, 1997].

Le critère déterminant les hypothèses à retenir est basé sur le calcul des probabilités $P(W)P(A|W)$, *i.e.*, sur le critère de la RAP (*cf.* équation (1.3)). En pratique, cet objectif se trouve modifié car les calculs de $P(A|W)$ et de $P(W)$ sont basés sur des hypothèses simplificatrices. Le MA suppose notamment que les probabilités $P(o_{t-1}|q_{t-1} = S_i)$ et $P(o_t|q_t = S_j)$ pour des vecteurs acoustiques o_{t-1} et o_t observés consécutivement sont indépendantes. Pour de nombreuses fenêtres d'étude du signal, du fait de la grande continuité acoustique entre des fenêtres successives, la probabilité de chaque sous-phone se trouve ainsi grandement sous-estimée en ignorant le contexte acoustique [Jurafsky et Martin, 2008]. Le poids entre $P(A|W)$ et $P(W)$ est donc changé par un *facteur d'échelle du ML*, noté α. De surcroît, l'utilisation de ML N-grammes tend à défavoriser les hypothèses contenant un plus grand nombre de mots. On a en effet :

$$P(w_1^n) = P(w_1^{n-1})P(w_n|w_{n-N+1}^{n-1}) \leq P(w_1^{n-1}) \tag{1.9}$$

puisque $P(w_n|w_{n-N+1}^{n-1}) \leq 1$. Ce phénomène est pris en compte par l'ajout d'une *pénalité d'insertion de mot* PIM. L'objectif du système de RAP devient ainsi :

$$\hat{W} = \arg\max_{W \in \mathcal{L}} \; P(A|W)\, P(W)^{\alpha}\, PIM^n \tag{1.10}$$

où α est le facteur d'échelle du ML par rapport au MA et n est le nombre de mots de W.

Lors des calculs, les log-probabilités sont généralement préférées aux probabilités dans la mesure où la multiplication de valeurs faibles par les processeurs se révèle imprécise. Par conséquent, l'égalité précédente est transformée comme suit :

$$\hat{W} = \arg\max_{W \in \mathcal{L}} \left[\log P(A|W) + \alpha \times \log P(W) + \gamma \times n\right] \tag{1.11}$$

en posant $\gamma = \log PIM$. Les coefficients α et γ sont déterminés empiriquement en cherchant à minimiser le nombre d'erreurs de transcription sur un corpus de développement, différent des corpus d'apprentissage utilisés pour fixer les paramètres du MA et du ML.

Lors du décodage, ces log-probabilités sont calculées efficacement pour les différentes hypothèses du graphe de mots à l'aide de l'algorithme de Viterbi [Viterbi, 1967] ou d'un décodage par pile, appelé aussi décodage A* [Hart *et al.*, 1968].

1.1.6 Sorties des systèmes de transcription

Comme dit précédemment, le rôle d'un système de RAP est de produire une transcription d'un signal sonore. Le résultat pourra donc être naturellement un texte. Il est de

20030418_0700_0800_inter_dga_spk1 : France Inter. *Franck Mathevon* : Le gouvernement fait sa pub. C'est aujourd'hui le vrai rendez-vous de la réforme des retraites, annonce-t-il. François Fillon et Jean-Paul Delevoye, les ministres des Affaires sociales et de la Fonction publique, dévoilent leur jeu aux syndicats aujourd'hui et mardi prochain, mais une partie des cartes a déjà été montrée avant-hier pour le secteur public et la partie de poker commence à déplaire sérieusement aux confédérations.
20030418_0700_0800_inter_dga_spk1 : france inter *Franck Mathevon* : le gouvernement fait sa plus basse et aujourd'hui le vrai rendez vous de la réforme des retraites annonce t il *Franck Mathevon* : françois fillon et jean paul delevoye le ministre des affaires sociales et de la fonction publique *Franck Mathevon* : dévoilant jour syndicats aujourd'hui mardi prochain *Franck Mathevon* : mais une partie des cartes déjà été montrées avant hier pour le secteur public *Franck Mathevon* : et la partie de poker comment se déplacer allemand aux confédérations

FIG. 1.4: *Extrait d'une transcription manuelle suivi de l'extrait correspondant d'une transcription produite automatiquement en utilisant un système de reconnaissance du locuteur*

plus possible de retenir plusieurs hypothèses pour un même groupe de souffle, certains mots mal reconnus dans la meilleure hypothèse proposée pouvant figurer correctement dans les suivantes ; la sortie pourra alors être une liste de $\mathcal{N}$ meilleures hypothèses, un graphe de mots ou encore un réseau de confusion.

Texte La sortie textuelle produite par un système de RAP possède des caractéristiques particulières par rapport à une transcription constituée manuellement. Selon les performances des systèmes de RAP (*cf.* section 1.4.2), le nombre de mots mal transcrits pourra être plus ou moins important. De plus, selon le vocabulaire utilisé, le résultat produit par certains systèmes de RAP peut ne contenir aucune majuscule, ce qui peut être problématique si l'on souhaite détecter automatiquement des noms propres. La segmentation est en outre effectuée en groupes de souffle et peut, dans certains cas, être bien différente de celle en phrases faite par un annotateur (*cf.* section 2.1.1), d'autant plus que les textes produits par un système de RAP ne contiennent pas de marque de ponctuations. Lorsqu'un système de reconnaissance du locuteur est utilisé au cours de la transcription, chaque groupe de souffle peut être toutefois accompagné de l'identité du locuteur courant. La figure 1.1.6 présente un extrait de transcription produite automatiquement.

Liste des $\mathcal{N}$ meilleures hypothèses Cette liste ordonne les hypothèses trouvées pour chaque groupe de souffle selon le score établi par un système de RAP, en fixant la

```
1: on a la lecture de paris
2: or a la lecture de paris
3: on à la victoire de paris
4: on a la lecture de paris
5: à la lecture a de paris
```

FIG. 1.5: *Exemple de liste des 5 meilleures hypothèses trouvées*

taille maximum $\mathcal{N}$ du nombre d'hypothèses[7]. Elle présente généralement de nombreuses redondances, ses éléments différant bien souvent par un seul mot (Fig. 1.5). Les successions de mots les plus probables sont déterminées à partir du graphe de mots construit par le décodeur en adaptant l'algorithme de Viterbi [Schwartz et Austin, 1991].

Graphes de mots Une sortie alternative à celle décrite précédemment est le graphe de mots. Ce graphe peut être une variante de celui établi par le décodeur. Les arcs sont valués par les probabilités établies par le MA et le ML, et représentent des hypothèses soit de mots prononcés, soit de pauses silencieuses (notées par « *sil* » sur la figure 1.6). Les nœuds sont quant à eux étiquetés par les instants supposés où un mot se termine et un autre débute. Puisque les informations concernant l'instant de prononciation des mots sont peu employées, le graphe est souvent compacté en supprimant les arcs et les nœuds qui représentent les mêmes hypothèses de succession de mots [Deshmukh *et al.*, 1999]. De surcroît, la taille du graphe de mots construit pouvant être considérable, il peut être utile de l'élaguer en fixant par exemple un nombre maximum de nœuds. Un autre critère d'élagage possible est le nombre $\mathcal{N}$ d'hypothèses retenues sur les successions de mots.

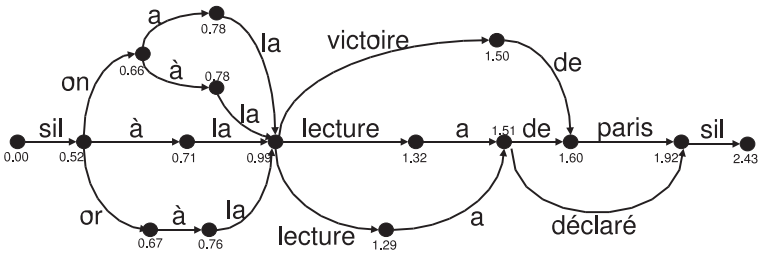

FIG. 1.6: *Exemple de graphe de mots (arcs non valués)*

Réseaux de confusion Cet autre type de sortie correspond à un compactage des graphes de mots, en conservant davantage d'informations que les $\mathcal{N}$ meilleures hypothèses. La construction des réseaux de confusion consiste à aligner les hypothèses de

[7] Le choix de $\mathcal{N}$ étant indépendant du paramètre N des ML N-grammes.

succession de mots, un peu comme on le ferait si l'on souhaitait aligner la transcription automatique avec la transcription de référence (*cf.* section 1.3). Ils sont également appelés « saucisses » car ils ont une forme de chapelet de saucisses, chacune étant associée à une position dans un groupe de souffle (Fig. 1.7). L'extraction des meilleures hypothèses se fait ici en tentant de minimiser les erreurs de transcription, et non en déterminant la succession de mots $\hat{W}$ possédant la plus grande probabilité $P(A|\hat{W})P(\hat{W})$, comme dans le cas des graphes de mots. Ceci est une propriété intéressante des réseaux de confusion dans la mesure où les systèmes de RAP cherchent plutôt à obtenir une transcription ayant le minimum d'erreurs qu'un texte ayant un bon « score » de probabilité. Les arcs du réseau de confusion correspondent à des hypothèses de mot à une position k donnée dans le groupe de souffle ; ils sont généralement valués par la probabilité *a posteriori* $P(w$ à la position $k)$ en sommant tout d'abord les probabilités de tous les chemins passant par w, puis en normalisant le résultat par la somme des probabilités de tous les autres mots placés en position k [Mangu *et al.*, 2000].

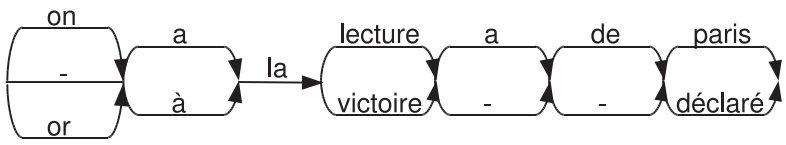

FIG. 1.7: *Exemple de réseaux de confusion (arcs non valués)*

1.1.7 Mesures de confiance

Dans certains contextes d'utilisation des systèmes de RAP, produire une transcription est insuffisant, surtout si la qualité des sorties obtenues n'est pas satisfaisante. Des applications, en particulier les systèmes automatiques de renseignement, requièrent en effet de savoir si les mots de la transcription sont corrects ou non. Dans le cas d'un système de réservation d'avion par exemple, il est ainsi préférable de savoir si la reconnaissance des paramètres du vol est sûre afin d'éviter de prendre une décision erronée. Pour ce faire, certains systèmes de RAP calculent un score de préférence entre 0 et 1, appelé *mesure de confiance*, indiquant la fiabilité des mots transcrits. Cet indice peut en outre être employé pour détecter les mots qui ne font pas partie du dictionnaire de prononciation [Asadi *et al.*, 1990] ou pour améliorer la qualité de la transcription [Wessel *et al.*, 2000]. Plusieurs méthodes sont envisageables pour l'obtenir automatiquement à partir d'informations comme la qualité acoustique, le nombre d'hypothèses alternatives ou la durée [Jiang, 2005]. Les probabilités *a posteriori* des mots sont notamment fréquemment utilisées pour le calculer [Wessel *et al.*, 2001].

À l'issue de la description de chacun des composants d'un système de RAP, il apparaît que le ML est primordial dans l'intégration de connaissances sur le langage. La section

suivante décrit plus en détail les ML N-grammes qui, bien que très utilisés, montrent certaines limitations pour prendre en compte certains phénomènes.

1.2 Limitations et extensions des modèles N-grammes

La section 1.1.4 nous a permis d'évoquer les modèles N-grammes. Cette modélisation, pour simpliste qu'elle soit, reste encore aujourd'hui à la base de la très grande majorité des ML actuels. Son principal avantage est qu'elle repose sur un processus gauche-droit, ce qui lui permet d'estimer les probabilités sans avoir vu tout le groupe de souffle et donc d'être utilisé par le décodeur au fur et à mesure que les mots sont reconnus. Cependant, ces modèles ont deux limitations principales qui sont l'éparsité des données et l'utilisation d'un historique de petite taille. Dans la suite de cette section, nous passons en revue chacune de ces restrictions en décrivant les solutions proposées dans le cadre des ML N-grammes.

1.2.1 Éparsité des données

Nous avons vu que pour limiter le nombre de paramètres, les ML N-grammes reposent sur l'approximation $P(w_i|w_1 \dots w_{i-1}) \approx P(w_i|w_{i-N+1} \dots w_{i-1})$ (*cf.* équation (1.7)). Cependant, ce mode de calcul nécessite encore de déterminer le nombre d'occurrences des $|\mathcal{V}|^N$ séquences de N mots possibles, appelées aussi des N-grammes ; si $|\mathcal{V}|$ devient conséquent, de nombreux N-grammes sont rares, voire absents du corpus d'apprentissage du ML, même pour une valeur réduite de N. Ce phénomène de la langue naturelle est expliqué par la loi de Zipf, ordinairement invoquée pour les mots simples pour justifier que la fréquence d'apparition d'un mot décroît rapidement avec son rang d'apparition, mais qui est observée de manière similaire pour les successions de mots [Ha *et al.*, 2002].

Lissage Une première modification possible des ML N-grammes visant à pallier les effets de l'éparsité des données repose sur des méthodes statistiques de lissage. Elle consiste à définir un facteur d'escompte (appelé aussi *discounting*) qui diminue le nombre d'occurrences, et donc la probabilité, pour les N-grammes observés dans le corpus d'apprentissage ; la masse de probabilité ainsi réduite est alors distribuée à la probabilité des N-grammes absents ou rares. Une méthode très simple de lissage consiste à additionner un facteur δ aux nombres d'occurrences observées pour chaque séquence de mots. Ce lissage, dit *additif*, redéfinit le calcul des probabilités par :

$$P(w_i|w_{i-N+1}^{i-1}) = \frac{\delta + C(w_{i-N+1}^{i})}{\delta \times |\mathcal{L}| + C(w_{i-N+1}^{i-1})} \quad . \tag{1.12}$$

Beaucoup d'autres algorithmes, parmi lesquels on peut citer le lissage de Good-Turing [Good, 1953], de Witten-Bell [Witten et Bell, 1991] ou celui de Kneser-Ney [Kneser et Ney, 1995]), reposent sur l'utilisation des N-grammes apparus une seule fois dans

le corpus d'apprentissage pour estimer les occurrences de ceux qui n'ont pas été observés. La technique de Good-Turing étend même ce principe en modifiant le nombre d'occurrences des mots apparus c fois en c^* à partir de celui des mots apparus $c+1$ fois :

$$c^* = (c+1)\frac{N_{c+1}}{N_c} \tag{1.13}$$

où N_c est le nombre de N-grammes qui ont été observés exactement c fois dans le corpus d'apprentissage.

D'après des tests effectués, la technique de lissage considérée actuellement comme la meilleure pour concevoir des ML N-grammes est une version modifiée de l'algorithme de Kneser-Ney [Chen et Goodman, 1998]. Elle s'inspire du lissage de Good-Turing mais utilise aussi l'interpolation linéaire et le repli, décrits ci-après.

Interpolation linéaire Une autre solution au problème de l'éparsité des données est de combiner linéairement les probabilités conditionnelles pour différentes tailles d'historique w_{i-N+1}^{i-1}. En remarquant que les nombres d'occurrences diminuent rapidement quand on augmente la grandeur N, cela consiste à évaluer $\hat{P}(w_i|w_{i-N+1}^{i-1})$ à partir des probabilités calculées pour les N-grammes d'ordre inférieur ou égal, *i.e.*, pour les k-grammes avec $k \leq N$. Dans le cas d'un modèle trigrammes par exemple, le calcul s'effectue à partir des probabilités unigrammes, bigrammes et trigrammes de la manière suivante :

$$\hat{P}(w_i|w_{i-2}w_{i-1}) = \lambda_1 P(w_i|w_{i-2}w_{i-1}) + \lambda_2 P(w_i|w_{i-1}) + \lambda_3 P(w_i) \tag{1.14}$$

avec :

$$\sum_{k=1}^{3} \lambda_k = 1 \tag{1.15}$$

de manière à ce que $\hat{P}$ demeure une probabilité. P est calculée au moyen de l'équation (1.6) utilisant les comptages. Les valeurs λ_k sont estimées de façon à maximiser la vraisemblance de $\hat{P}$ sur un corpus de développement, différent du corpus d'apprentissage.

Repli Une autre stratégie pour déterminer les probabilités à partir de celles estimées à des ordres inférieurs est celle du repli (appelé aussi *backoff*). Elle consiste à préférer les modèles plus complexes en premier lieu, avant d'envisager des plus simples quand on dispose de peu d'informations par rapport à l'historique courant [Katz, 1987]. Elle s'exprime sous la forme :

$$\hat{P}(w_i|w_{i-N+1}^{i-1}) = \begin{cases} P^*(w_i|w_{i-N+1}^{i-1}) & \text{si } C(w_{i-N+1}^{i}) > 0 \\ \alpha(w_{i-N+1}^{i-1}) \times \hat{P}(w_i|w_{i-N+2}^{i-1}) & \text{sinon} \end{cases} \tag{1.16}$$

De même que pour les λ_i de l'équation (1.14), les coefficients α sont calculés pour que $\hat{P}$ soit une probabilité. Le symbole $*$ apposé au P sert à indiquer que P^* est souvent obtenu à partir d'un procédé de lissage.

Modèle N-classes D'autres solutions consistent à définir de nouveaux modèles, inspirés des ML N-grammes. Les modèles N-classes ont ainsi souvent été envisagés. Leur conception repose sur l'existence dans la langue de catégories ou de classes qui apparaissent souvent dans les mêmes contextes. Ces classes peuvent être construites à l'aide de méthodes de classification automatiques, en cherchant par exemple à maximiser l'information mutuelle entre les classes [Brown *et al.*, 1992 ; Jardino, 1996] ou à minimiser la perplexité[8] [Kneser et Ney, 1993] ou encore en rassemblant dans une classe donnée les mots qui apparaissent avec les mêmes mots adjacents [Farhat *et al.*, 1996]. Elles peuvent également être déterminées *a priori* à partir de connaissances sur $\mathcal{L}$ (*cf.* chapitre 2).

Dans les modèles N-classes, si $\mathcal{C}_i$ représente l'ensemble des classes c_i auxquelles peut appartenir un mot w_i, le calcul des probabilités se fait de la manière suivante :

$$P(w_1^n) = \sum_{c_1 \in \mathcal{C}_1 \ldots c_n \in \mathcal{C}_n} P(w_1^n, c_1^n) \tag{1.17}$$

$$= \sum_{c_1 \in \mathcal{C}_1 \ldots c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i | w_1^{i-1}, c_1^i) P(c_i | w_1^{i-1}, c_1^{i-1}) \tag{1.18}$$

où $P(w_i|w_1^{i-1}, c_1^i)$ est appelée la *probabilité lexicale* et $P(c_i|w_1^{i-1}, c_1^{i-1})$ la *probabilité contextuelle*. En supposant que la probabilité de w_i dépend uniquement des classes c_i auxquelles il peut appartenir dans le groupe de souffle, on obtient :

$$P(w_1^n) \approx \sum_{c_1 \in \mathcal{C}_1 \ldots c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i | w_1^{i-1}, c_1^{i-1}) \ . \tag{1.19}$$

De manière similaire aux modèles N-grammes de mots qui ne prennent en compte que les $N-1$ mots précédents dans le calcul, $P(c_i|w_1^{i-1}, c_1^{i-1})$ est approximé en ne considérant que les $N-1$ classes attribuées précédemment :

$$P(w_1^n) \approx \sum_{c_1 \in \mathcal{C}_1 \ldots c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i | c_{i-N+1}^{i-1}) \ . \tag{1.20}$$

L'intérêt principal des modèles N-classes est de réduire considérablement le nombre d'événements possibles par rapport aux modèles N-grammes puisque le nombre total de classes est généralement beaucoup plus petit que la taille du vocabulaire du système de RAP. Ceci permet de limiter le recours aux techniques de lissage, qui introduisent des approximations lors du calcul des probabilités. Toutefois, l'affectation à une classe statistique se révèle problématique pour les mots absents des corpus et peu fiable pour les mots rares [Rosenfeld, 2000a].

[8] Définie à la section 1.3.1.

1.2.2 Historique de petite taille

Si l'on restreint le ML à un processus gauche-droit, nous avons vu (*cf.* équation (1.4)) que le calcul de la probabilité d'une séquence de mots pouvait être écrite :

$$P(w_1^n) = P(w_1) \prod_{i=2}^{n} P(w_i|w_1^{i-1}) \ . \tag{1.21}$$

Afin de restreindre le nombre de paramètres à estimer et de réduire les problèmes d'éparsité des données, une approximation est réalisée en conditionnant le mot w_i par une classe d'équivalence $\phi(w_1^{i-1})$ de l'historique [Jelinek, 1998] :

$$P(w_1^n) \approx P(w_1) \prod_{i=2}^{n} P(w_i|\phi(w_1^{i-1})) \ . \tag{1.22}$$

Dans le cas des ML N-grammes, $\phi(w_1^{i-1}) = w_{i-N+1}^{i-1}$, en utilisant rarement des valeurs de N supérieures à 4. De nombreuses dépendances linguistiques excédant cette taille se trouvent ainsi ignorées par le modèle. Dans l'exemple suivant : « *les oiseaux sur l'arbre chantaient* », un modèle quadrigrammes n'aura pas les éléments utiles, dans son historique de trois mots, pour prédire l'accord du verbe « *chantaient* », du fait de l'insertion du complément de lieu entre le sujet et le verbe. Divers modèles statistiques ont été conçus pour prendre en compte l'influence de phénomènes à plus longue distance.

Historique à taille variable Pour accroître la taille du contexte pris en compte dans le calcul des probabilités, sans augmenter de manière exponentielle le nombre de paramètres à estimer, certains ML utilisent des historiques dont la longueur dépend des mots w_1^{i-1} reconnus auparavant. Plusieurs fonctions d'équivalences ϕ ont été envisagées. Une première solution consiste à considérer que $\phi(w_{i-N+1}^{i-1}) = \phi(w_{i-N+2}^{i-1})$ si w_{i-N+1}^{i} a été peu observé dans le corpus d'apprentissage ou si la distribution de probabilité $P(w) = P(w|w_{i-N+1}^{i-1})$ est similaire à $P(w) = P(w|w_{i-N+2}^{i-1})$ [Kneser, 1996 ; Riccardi *et al.*, 1996 ; Niesler et Woodland, 1996b ; Siu et Ostendorf, 2000]. Une deuxième méthode revient à avoir des historiques « à trous » en ayant par exemple $\phi(w_1^{i-1}) = w_{i-4}w_{i-3}w_{i-1}$ [Goodman, 2001 ; Schukat-Talamazzini *et al.*, 1995], ce qui permet d'exclure notamment des termes peu informatifs comme des répétitions de mots (*cf.* section 2.1.3) [Siu et Ostendorf, 2000]. Un troisième type de ML, appelé modèles *multigrammes*, considère l'historique non pas comme une succession de mots mais comme une suite de groupes de mots [Bimbot *et al.*, 1995 ; Deligne et Bimbot, 1995].

Historique différent d'une séquence de mots Le fait de considérer l'historique comme un « sac de mots » permet de l'étendre jusqu'à plusieurs centaines de mots puisque les événements, réduits à des unigrammes, sont beaucoup moins nombreux. Les modèles à base de cache[9] utilisent cette méthode, en supposant qu'un mot observé récemment risque à nouveau d'être reprononcé [Kuhn et De Mori, 1990]. Les modèles

[9] Appelée ainsi en référence au principe de fonctionnement de la mémoire cache des processeurs.

dits à amorces, appelés aussi *triggers*, étendent l'influence entre deux mots distants à des mots différents comme « *orchestre* » et « *concert* » [Tillmann et Ney, 1997 ; Rosenfeld, 1997].

On le voit donc, les deux problèmes principaux liés à l'utilisation des ML N-grammes sont l'éparsité des données et une limitation de la taille de l'historique à une valeur de N faible. Malgré leur nombre, les différentes variantes statistiques proposées sont généralement reconnues comme n'apportant pas de gain significatif au niveau de la qualité de la transcription, eu égard à leur complexité algorithmique ou à leur nombre de paramètres, par rapport à un ML N-grammes utilisant des techniques de lissage et de repli [Goodman, 2001]. Certains ML peuvent néanmoins être utilisés en post-traitement, *i.e.*, quand le nombre d'hypothèses a été réduit après élagage du graphe de mots construit par le décodeur, pour apporter des informations complémentaires à celles dont dispose un ML N-grammes. Une alternative aux modèles purement statistiques que nous avons mentionnés repose sur la prise en compte de connaissances linguistiques. Le chapitre 2 est consacré à la description de ce point de vue.

Afin de comparer les différents modèles proposés, plusieurs mesures de leur qualité ont été envisagées ; la section suivante présente les principales.

1.3 Méthodes d'évaluation

Nous exposons ici les mesures d'évaluation qui sont couramment utilisées en RAP, en débutant par celles qui sont employées spécifiquement pour comparer les performances des ML, avant de présenter celles mesurant la qualité globale d'un système de transcription.

1.3.1 Évaluation des modèles de langage

Il existe deux méthodologies pour évaluer un ML : soit on ne fait pas d'hypothèse sur le contexte applicatif, les ML pouvant servir dans bien d'autres cadres que celui de la reconnaissance de la parole[10], soit on utilise les mesures de qualité de l'ensemble du système de transcription, en laissant tous les composants inchangés à l'exception du ML.

La mesure la plus employée pour connaître l'efficacité du ML à modéliser un langage est la *perplexité*. Elle se détermine sur un corpus de test $T = w_1^n$ représentatif du langage, à partir de la probabilité de T calculée par le ML et normalisée par le nombre de mots :

$$PP_T = \sqrt[n]{\frac{1}{\hat{P}(w_1 \dots w_n)}} \ . \tag{1.23}$$

D'après la forme de cette égalité, moins la perplexité est grande, meilleur est le ML pour prédire les mots du corpus de test. Cette mesure est directement liée à celle de

[10] *E.g.* la reconnaissance de texte manuscrit ou la recherche d'information.

l'entropie croisée, calculée sur T, entre la distribution de probabilité $\hat{P}$ de $\mathcal{L}$ estimée par le ML et la distribution P réelle de $\mathcal{L}$:

$$H_T(P, \hat{P}) = -\frac{1}{n} \log_2 \hat{P}(w_1 \dots w_n) = \log_2 PP_T \ . \tag{1.24}$$

L'entropie croisée est une mesure issue de la théorie de l'information permettant de mesurer la qualité du ML pour estimer P. De la même manière que la perplexité, plus l'entropie croisée est faible, plus le ML est performant.

Bien que très utilisé pour la comparaison des ML, le critère de la perplexité est sujet à de nombreuses critiques, une baisse importante ne correspondant pas forcément à une baisse de même ordre du nombre d'erreurs dans la transcription, si on intègre le ML dans un système de RAP ; seules des réductions de la perplexité d'au moins 10-20 % semblent être vraiment significatives [Rosenfeld, 2000b]. Son principal inconvénient est qu'il favorise les modèles accordant une plus grande probabilité aux mots présents dans le corpus de test, et ignore la manière avec laquelle sont distribuées les probabilités aux autres mots. Or, ces mots peuvent conduire, lors du décodage par le système de RAP, à des probabilités plus élevées que celles obtenues pour les mots corrects [Clarkson et Robinson, 1999]. Certains auteurs préconisent donc d'utiliser plutôt l'entropie croisée, qui semble plus corrélée avec le taux d'erreur [Goodman, 2001]. D'autres privilégient des mesures alternatives, inspirées du jeu de Shannon [Shannon, 1951 ; Bimbot *et al.*, 1997 ; Alain et Boëffard, 2006].

1.3.2 Évaluation des systèmes de transcription

Contrairement à l'évaluation des ML pour laquelle il n'y a pas de véritable consensus sur la mesure à choisir, la qualité d'une transcription est presque systématiquement caractérisée par son *taux d'erreurs sur les mots* (notée WER pour *word error rate*). Ce taux est mesuré en alignant le texte produit par un système de RAP (HYP) avec la transcription de référence (REF), obtenue généralement manuellement. L'alignement est réalisé en minimisant la distance d'édition [Levenshtein, 1966] entre REF et HYP au niveau de chaque groupe de souffle. Trois types d'erreur sont distingués (Fig. 1.8) :

- les insertions, repérées par des I, correspondant à un ajout d'un mot de la transcription automatique par rapport à la transcription de référence,
- les suppressions, marquées par des D (pour *deletion*), associées à un mot manquant dans la transcription automatique,
- les substitutions, représentées par des S, indiquant un remplacement d'un mot de la transcription de référence par un autre mot présent dans la transcription automatique.

Le WER est établi en calculant la distance minimale d'édition entre la référence et l'hypothèse pour les groupes de souffle d'un ensemble de test :

$$\text{WER} = 100 \times \frac{\text{nb d'insertions} + \text{nb de suppressions} + \text{nb de substitutions}}{\text{nb de mots dans la transcription de référence}} \ . \tag{1.25}$$

Un autre indicatif peut être utilisé en complément du WER ; il s'agit du *taux d'erreurs sur les phrases* (noté SER pour *sentence error rate*), donnant la proportion de

```
REF: il AURA ALORS  face à lui une fronde syndicale *** UNIE
HYP: il **** VALEUR face à lui une fronde syndicale EST PUNI

         D    S                                        I   S
```

FIG. 1.8: *Alignement d'une transcription automatique (HYP) et d'une transcription de référence (REF)*

groupes de souffle ayant au moins une erreur de transcription :

$$\text{SER} = 100 \times \frac{\text{nb de groupes de souffle avec au moins un mot erroné}}{\text{nb total de groupes de souffle}} \ . \qquad (1.26)$$

Il est parfois reproché au WER de ne pas prendre en compte la valeur informative de chaque mot dans son calcul puisque une erreur commise sur un mot de contenu comme « *grève* » a autant de poids que sur un mot grammatical tel que « *de* ». Dans certaines applications, notamment celles intégrant un système de compréhension de la parole, d'autres métriques peuvent être préférées [Wang *et al.*, 2003b].

Après une présentation des principales techniques d'évaluation des systèmes de RAP, nous exposons dans la section suivante les problèmes posés lors de la transcription suivant la complexité du document à analyser, en indiquant les performances obtenues par les meilleurs systèmes.

1.4 Difficultés de la transcription

La RAP est généralement reconnue comme étant difficile, en particulier quand on cherche à traiter des documents sonores très divers. Cette section présente les problèmes principaux rencontrés lorsque l'on transcrit de la parole, avant d'indiquer quelles sont les performances généralement obtenues par des systèmes de RAP.

1.4.1 Une grande variabilité

Les difficultés de la RAP sont dues pour une grande part à la diversité des signaux à traiter. Cette variabilité intervient au niveau des propriétés acoustiques du signal et au niveau des propriétés du langage $\mathcal{L}$ du type de documents que l'on souhaite transcrire.

Au niveau acoustique La parole produite pour une même phrase prononcée peut tout d'abord varier d'un individu à un autre. Outre le fait que chaque individu possède une voix qui lui est propre, on rencontre d'importantes différences telles que les variations homme/femme, le régionalisme ou encore les difficultés de prononciation rencontrées par des locuteurs non natifs. Cette variabilité est qualifiée d'*inter-locuteurs*. Il existe également une variabilité, dite *intra-locuteur*, correspondant à une modification de la parole produite par un même individu. Cette variabilité peut concerner aussi bien

les caractéristiques de la voix, dans les cas de rhumes ou d'un état émotionnel, ou bien la qualité d'élocution, selon que la parole intervient lors d'un discours formel ou d'un dialogue spontané. En sus des variabilités au niveau de la parole prononcée par le locuteur, les conditions d'enregistrement peuvent dégrader le signal qui sera traité par le système de RAP. Il peut par exemple y avoir une modification de la qualité du signal, notamment si celui-ci doit transiter par un canal de communication qui a une bande passante limitée, comme une ligne téléphonique. De même, l'environnement acoustique peut être disparate. Le bruit de fond peut être plus ou moins important et de natures diverses (musique, paroles d'autres locuteurs, parasites du micro, bruits de bouche...).

Cette variabilité contraint le MA à devoir être robuste vis-à-vis du locuteur et de l'environnement. Pour transcrire des signaux spécifiques à un locuteur comme cela est le cas des dictées vocales, l'adaptation est réalisée en ayant un corpus d'apprentissage pour chaque nouvel utilisateur. Pour décoder la parole présente dans des émissions de radio ou de télévision — où certaines personnes absentes du corpus d'apprentissage du MA interviennent — des mécanismes statistiques permettent une adaptation à un nouveau locuteur [Woodland, 2001].

Au niveau du langage On retrouve une variabilité similaire au niveau du lexique ou des tournures employés dans les documents à analyser. Suivant la maîtrise de la langue par le locuteur ou selon que la conversation ait lieu dans un cadre formel ou courant, la manière de s'exprimer pourra différer. De même, selon les thèmes abordés, on peut observer une grande variation d'emplois des mots. La taille du vocabulaire à envisager selon le type de documents que l'on souhaite transcrire est, quant à elle, un facteur-clé influençant la qualité des résultats puisqu'elle influe directement sur la taille de l'espace de recherche à envisager lors du décodage. Si le lexique peut être très réduit (moins de 100 mots) pour naviguer dans un menu, les systèmes de RAP conçus pour transcrire des émissions d'information diffusées à la radio ou à la télévision avoisinent les 65 000, voire les 200 000 mots, ce qui demeure malgré tout insuffisant pour contenir tous les mots rencontrés, notamment les noms propres qui apparaissent au gré de l'actualité.

1.4.2 Performances des systèmes de transcription

Les performances des systèmes de transcription progressent régulièrement grâce à la conception d'algorithmes plus performants mais aussi avec l'augmentation des capacités de mémoire et de calcul des ordinateurs. Ceci conduit à envisager des documents de plus en plus complexes à transcrire puisque la communauté de la RAP s'est intéressée dans les années 50 à la reconnaissance de chiffres, pour ensuite étudier la lecture de journaux, les émissions d'actualité et dorénavant les échanges intervenant lors de réunions. Les campagnes d'évaluation conduites par le NIST depuis 1987 témoignent de cette évolution [Pallett, 2003].

Le WER des transcriptions dépend de la difficulté de la tâche envisagée. Ainsi peut-on espérer un WER inférieur à 1 % lorsque les mots à reconnaître sont prononcés de manière isolée, parmi un vocabulaire restreint. Quand la parole devient continue, l'espace des hypothèses à envisager augmente grandement puisqu'il faut déterminer les

frontières entre les mots, ce qui complique le décodage. Cependant, les meilleurs systèmes actuels parviennent à des WER voisins de 6 % pour reconnaître le *Wall Street Journal* lu, 10 % pour les émissions d'information et 20 % pour les conversations téléphoniques spontanées. Ces performances sont obtenues pour des documents sonores de langue anglaise mais elles restent voisines pour les autres langues ; pour le français, pour lequel il y a un peu moins de corpus disponibles et qui est une langue à haut taux de flexion, on peut ainsi obtenir un WER inférieur à 12 % pour transcrire des émissions d'information [Gauvain *et al.*, 2005]. La comparaison des taux d'erreur obtenus pour différentes tâches montre clairement que la difficulté dépend fortement de la spontanéité des documents produits. On obtient en effet des résultats supérieurs pour des journaux lus, qui sont au départ des textes écrits, que pour des conversations de la vie courante, contenant de nombreux phénomènes d'hésitation.

1.5 Bilan

Ce chapitre décrivant les principes de fonctionnement des systèmes de RAP a mis en exergue les différentes étapes nécessaires au décodage de la parole. Les méthodes généralement utilisées pour faire la relation entre le son et la séquence des mots prononcés sont essentiellement statistiques. Malgré les difficultés de la transcription dues pour une grande part à la variabilité des documents à traiter, la RAP est devenue suffisamment mature pour être utilisée dans des domaines limités. Toutefois, en décrivant plus particulièrement les modèles de langage, nous avons pu constater que les modèles N-grammes, bien que très utilisés en pratique, sont souvent critiqués sur deux points. D'une part, ils sont basés sur une hypothèse un peu simpliste et arbitraire qui est d'examiner les $N-1$ mots précédents pour prédire le mot courant. D'autre part, malgré des techniques de lissage perfectionnées, les calculs de probabilité réalisés pour des événements rares voire absents du corpus sont imprécis. Les performances des systèmes actuels semblent donc améliorables en recourant à des techniques supplémentaires qui n'auraient pas ces inconvénients.

Une des possibilités envisagées pour pallier les manques des méthodes statistiques pour modéliser certains phénomènes consiste à utiliser davantage d'informations sur la langue. Le chapitre suivant est dédié à un état de l'art sur les techniques employées pour introduire des connaissances linguistiques au sein des ML.

Chapitre 2

Connaissances linguistiques en reconnaissance de la parole : un état de l'art

Comme nous l'avons vu en section 1.1, les systèmes de RAP adoptent une modélisation hiérarchique pour décoder la parole. Le MA identifie, à partir des caractéristiques extraites du signal, des phones puis des mots ; le ML est quant à lui chargé de reconnaître les successions de mots les plus probables pour un groupe de souffle. Ce mode de fonctionnement présente des similarités avec celui de l'être humain qui, pour interpréter le sens véhiculé par la parole, identifie successivement des phones, des syllabes, des mots et des « phrases » [Allen, 1994]. Cependant, à la différence de l'être humain qui utilise différents niveaux de connaissances linguistiques, les informations sur le langage dans les systèmes de RAP se limitent essentiellement à la connaissance des phones et des phonèmes d'une langue, à la réalisation d'un dictionnaire de prononciation et à l'apprentissage de ML sur des corpus. Les ML notamment n'exploitent pas la richesse des informations liées à la langue et assimilent bien souvent les mots à de simples symboles.

Une justification de l'utilisation de connaissances linguistiques en RAP est qu'il semble de plus en plus difficile d'améliorer les modèles N-grammes en étendant la taille du corpus d'apprentissage. En effet, même si la collecte de nombreux textes de natures différentes permet de transcrire des documents issus de domaines de plus en plus ouverts depuis les débuts d'utilisation de modèles statistiques, il semble, selon une estimation informelle d'IBM, que les performances des modèles bigrammes n'enregistrent plus de gain important au-delà de quelques centaines de millions de mots, tandis que les modèles trigrammes saturent à partir de quelques milliards de mots. Or, dans plusieurs domaines d'application de systèmes de RAP, de tels volumes de données ont déjà été collectés [Rosenfeld, 2000b]. Une autre étude menée sur une application du TAL — la désambiguïsation d'un mot parmi un ensemble de mots souvent confondus, *e.g.* « *to* », « *too* » et « *two* » en anglais — conduit à une conclusion contradictoire sur l'intérêt d'accroître la taille d'un corpus d'apprentissage pour améliorer la performance des mé-

thodes statistiques, les auteurs montrant qu'il n'existe pas de plateau de saturation des résultats quand on continue à augmenter le nombre de données utilisées [Banko et Brill, 2001]. Toutefois, malgré ces résultats encourageants, la constitution de corpus spécifiques à chaque type de document à transcrire demeure nécessaire pour améliorer véritablement le WER [Rosenfeld, 2000b]. Or, la collecte de données pertinentes et en grand volume reste toujours une opération coûteuse et de longue haleine[1]. Pour toutes ces raisons, il apparaît ainsi nécessaire de recourir à d'autres connaissances pour permettre d'améliorer la qualité des modèles proposés.

Une expérience menée par Brill *et al.* corrobore l'hypothèse d'un apport de la linguistique à l'amélioration de la reconnaissance de la parole. Elle a consisté à analyser les ressources qu'un être humain utiliserait pour corriger des transcriptions automatiques [Brill *et al.*, 1998]. Trois corpus de parole ont été à cette fin traités par un système de RAP pour produire les listes des dix meilleures hypothèses (*cf.* section 1.1.6) associées à chaque groupe de souffle. Il a alors été demandé à des sujets humains de sélectionner parmi chacune de ces listes l'hypothèse qui leur semblait la plus juste. Les choix effectués par ces personnes se sont souvent révélés judicieux puisqu'ils ont permis une nette diminution du taux d'erreur par rapport à la meilleure hypothèse désignée par le système de RAP pour chacun des groupes de souffle. Un questionnaire sur les connaissances utilisées pour faire leur sélection a montré que les humains se basaient principalement sur des informations linguistiques telles que l'emploi correct des prépositions et des déterminants, les accords en genre et en nombre, l'examen du temps pour les verbes, la connaissance de syntagmes idiomatiques ou encore l'analyse de la structure des hypothèses proposées.

Ce chapitre se propose de faire un état de l'art sur l'utilisation de connaissances linguistiques pour améliorer les performances de systèmes de RAP[2]. D'une part, l'intervention de ce type d'informations suppose de les acquérir en examinant les données manipulées par le système de transcription. Pour cela, le domaine du traitement automatique des langues (TAL) fournit un ensemble de techniques pour analyser automatiquement du texte exprimé en langues naturelles, et la prise en compte de nouvelles sources de connaissances linguistiques en RAP peut ainsi être vue comme une des applications de l'utilisation conjointe de la RAP et du TAL. D'autre part, l'emploi des nouvelles connaissances pour améliorer la transcription nécessite de les inclure dans le processus de la RAP. Ceci peut demander d'adapter les méthodes usuelles de décodage.

Nous débutons par une présentation des caractéristiques de la langue parlée. L'emploi sur des productions d'un système de RAP de techniques du TAL, appliquées ordinairement sur des textes « bien écrits », ne va en effet pas de soi et se heurte à des difficultés. Nous poursuivons en exposant quels types de connaissances linguistiques peuvent être exploitées pour améliorer la qualité d'une transcription, comment elles sont obtenues par des techniques du TAL puis utilisées, et à quels résultats elles conduisent.

[1] L'utilisation des ressources du Web peut constituer une solution à ce problème comme nous le verrons au chapitre 6.

[2] Une version étendue de la section 2.2 figure dans le rapport interne [Huet *et al.*, 2006b].

Nous précisons enfin à quel niveau du décodage peuvent être prises en compte ces informations sur la langue.

2.1 Caractéristiques de la langue parlée

Si l'on excepte le cas particulier des signaux enregistrés lors de la lecture à haute voix de textes écrits, les documents traités par les systèmes de RAP relèvent du domaine de la langue parlée. Les corpus oraux qui ont pourtant certaines spécificités par rapport aux textes écrits, comme les romans ou les articles de journaux, ont globalement été peu étudiés par la communauté du TAL. Cette constatation a deux explications principales [Véronis, 2004]. D'une part, le mode de pensée de la culture occidentale tend à privilégier l'étude de l'écrit ; alors que les lois, les contrats, les textes religieux fondamentaux sont des documents écrits, l'oral ne nécessite pas un apprentissage à l'école aussi formel que l'écrit. D'autre part, peu de corpus oraux transcrits manuellement sont actuellement disponibles. Ainsi, si le *British National Corpus* (BNC) contient 100 millions d'occurrences pour l'anglais, tandis que la base *Frantext* comporte 210 millions d'occurrences pour le français, le plus grand corpus oral, à savoir la partie orale du BNC, est constitué de 10 millions d'occurrences. Parmi les corpus oraux, on peut citer pour le français de Belgique le corpus Valibel comptant près de 4 millions d'occurrences et pour le français de l'hexagone, le corpus Corpaix comptant environ 2,5 millions d'occurrences [Campione *et al.*, 2005], et le *Corpus de référence du français parlé* comptant 440 000 occurrences [Équipe DELIC, 2004]. Le faible volume de données est expliqué par la difficulté d'établir des corpus oraux puisqu'il a été estimé qu'environ 40 heures de travail sont nécessaires pour faire une transcription orthographique précise d'une heure de parole, et ce, avec de bonnes conditions d'enregistrement et sans alternance de locuteurs [Campione, 2001].

Parmi les caractéristiques de la langue parlée, certaines viennent compliquer la transcription de la parole et l'acquisition automatique de connaissances linguistiques. Après avoir précisé ce qui définit la langue parlée par rapport à la langue écrite, cette section présente les principales caractéristiques de cette langue parlée du point de vue de son vocabulaire et de sa syntaxe. Nous terminons en décrivant les perturbations liées aux phénomènes d'hésitation, souvent présents à l'oral.

2.1.1 Langue parlée et langue écrite

La *langue parlée* est définie comme étant ce qui est prononcé par des locuteurs à l'oral et fait donc appel à la voix et à l'oreille. La *langue écrite* représente quant à elle ce qui s'écrit et implique donc l'usage de la main et des yeux. Les langues parlée et écrite exploitent ainsi des canaux différents qui ont des contraintes importantes sur leur mode de production [Melis, 2000]. Le canal oral est ainsi caractérisé par une certaine linéarité lors de son émission et de sa réception, même s'il peut être possible d'effectuer des retours en arrière à l'aide de dispositifs d'enregistrement. La production orale ne possède pas en outre un aussi haut niveau de structuration que l'écrit puisque seules la notion de *tours de parole*, définis comme étant les suites de mots prononcés par un locuteur

avant qu'il ne cède la parole, ou encore celle de *groupes de souffle* sont des unités facilement identifiables. La *prosodie*, recouvrant des informations sur l'intonation, l'accent, les pauses ou même le débit, joue également un rôle important dans la structuration mais cet indice est beaucoup plus difficile à analyser que ne le sont les ponctuations à l'écrit ; les pauses silencieuses se révèlent par exemple imprécises pour détecter des ponctuations dans un signal correspondant à un texte lu [Campione et Véronis, 2002]. Des méthodes ont été élaborées pour obtenir automatiquement une segmentation plus linguistique que celles par groupes de souffle, ceux-ci étant uniquement basés sur des indices acoustiques et n'ayant pas la même cohérence que des phrases. Cependant, cela reste une tâche difficile puisque les meilleurs systèmes détectent les fins de phrases avec des taux d'erreurs variant entre 30 % et 50 % selon le type de discours à analyser [Liu *et al.*, 2004].

Le canal écrit quant à lui fait intervenir deux dimensions : la largeur et la hauteur, facilitant ainsi les retours à des éléments antérieurs. Ceci offre la possibilité d'avoir une structuration plus importante puisqu'un texte peut être segmenté en chapitres, en sections, en paragraphes et en phrases. Ce découpage est lui-même choisi par celui qui écrit et est plus informatif sur l'alternance des points abordés que ne le sont par exemple les tours de parole.

En sus du format, l'oral et l'écrit s'opposent aussi sur leur contexte d'utilisation. Ainsi l'oral exclut-il généralement toute forme de préparation préalable ; il permet davantage d'interactivités entre les interlocuteurs, au risque même d'avoir des chevauchements entre des mots prononcés par des personnes différentes. Le texte a plutôt tendance à être produit en différé et peut donc être reformulé avant d'être livré aux destinataires. Les récepteurs ne sont pas au contact du scripteur, ce qui contraint celui-ci à proposer un texte désambiguïsé, informatif et structuré. Ces différences d'utilisation entraînent des conséquences importantes sur le style plus formel à l'écrit qu'à l'oral. Avec l'apparition de nouvelles formes de documents tels que les courriels ou les *blogs*, l'opposition entre langue parlée et langue orale tend cependant à être moins systématique et on assiste à une hybridation des codes.

D'une manière générale, les différences observées entre écrit et oral dépendent toutefois du contexte d'utilisation. Le discours télévisé du chef de l'état à la nation sera par exemple préparé et n'autorisera pas l'intervention des récepteurs, tandis que les conversations de la vie courante seront plus sujettes à des marques d'hésitation et sont très interactives. On distingue trois grandes catégories de documents étudiés en RAP : la parole *lue*, *préparée* ou encore *spontanée*. La parole lue est associée à la lecture à haute voix de textes, typiquement des articles de journaux. La parole spontanée correspond à des situations de la vie quotidienne où le locuteur parle sans avoir préparé au préalable ce qu'il allait dire et en s'adaptant constamment à ses interlocuteurs. La parole préparée est une forme intermédiaire par rapport aux deux précédentes et est plus délicate à définir. Elle désigne des prises de parole où le locuteur a réfléchi auparavant aux idées qu'il souhaite transmettre ; il peut même disposer d'un texte contenant une partie de son discours, qu'il n'ira pas toutefois jusqu'à lire mot à mot pour pouvoir interagir avec les interlocuteurs. Les questions posées par un journaliste et les entretiens donnés par un homme politique sont deux exemples de production de parole préparée.

Après avoir défini ce qu'était la langue parlée, nous nous proposons de dégager des phénomènes revenant plus fréquemment à l'oral qu'à l'écrit et qui complexifient une analyse par des techniques automatiques.

2.1.2 Vocabulaire et syntaxe

Le registre de la langue parlée est généralement moins soutenu dans la mesure où elle est souvent employée dans des cadres moins formels, ce qui fait qu'elle se fonde en général sur un ensemble familier de mots, plus restreint que celui de la langue écrite. L'utilisation des catégories grammaticales semble également différer puisque la comparaison de deux corpus, l'un correspondant à des journaux écrits et l'autre à des émissions journalistiques radiophoniques et télévisées, a montré que l'oral avait tendance à contenir plus de pronoms et moins de noms qu'à l'écrit [Gendner et Adda-Decker, 2002]. La langue parlée est également marquée par l'introduction de « petits » mots, appelés ligateurs, marqueurs de discours ou encore inserts, comme par exemple « *quoi* », « *bon* », « *donc* », « *enfin* » ou « *genre* ».

Au niveau de la syntaxe, il existe des tournures propres à l'oral. Si en français le « *pas* » est parfois facultatif à l'écrit dans la négation « *ne ... pas* », c'est le « *ne* » qui le devient à l'oral. Le « *il* » prend également à l'oral un caractère facultatif dans les formules « *il faut* » et « *il y a* ». On peut aussi citer l'invariabilité du « *c'est* » dans des expressions telles que « *c'est les voisins qui sonnaient* ». Ces expressions sont dues au contexte d'utilisation de la langue orale qui favorise l'interactivité et impose donc un temps limité de formulation des idées. Pour la même raison, et même si cela est loin d'être systématique, les accords du participe passé seront moins respectés à l'oral qu'à l'écrit [Melis, 2000].

Il semble également que l'ordre des mots soit un peu plus souple dans la langue parlée qu'à l'écrit, bien que les langues rigides sur l'ordre des mots comme le français ou l'anglais le soient encore à l'oral [Antoine et Goulian, 2001]. Les éléments régis par le verbe peuvent, à l'oral, se placer avant le groupe sujet-verbe, comme dans l'exemple « *les haricots j'aime pas* ». On observe aussi davantage de clivées (« *c'est le coiffeur qui est content* »), de pseudos-clivées (« *ce qui l'intéresse c'est le pognon* »), de doubles marquages (« *moi j'en ai jamais vu en Suisse des immeubles* ») ou encore de dislocations (« *j'ai choisi la bleue de robe* ») [Blanche-Benveniste, 1990].

Ces phénomènes ne justifient pas toutefois de proposer une grammaire spéciale pour la langue parlée ; pour le français, Blanche-Benveniste indique ainsi que la syntaxe de l'oral ne diffère en rien de celle de l'écrit, sauf en termes de proportions [Blanche-Benveniste, 1990 ; Benzitoun *et al.*, 2004]. Toutefois, ils expliquent l'importance d'avoir à disposition des corpus oraux pour l'apprentissage des ML. En pratique, les données recueillies pour la langue parlée n'atteignent pas une taille suffisante et on utilise généralement un corpus supplémentaire issu de la langue écrite.

2.1.3 Phénomènes d'hésitation

La fréquence d'apparition des marques d'hésitation est très variable selon le locuteur et le contexte d'élocution. Toutefois, elles concernent toutes les productions de l'oral et sont particulièrement fréquentes à l'oral spontané puisqu'il est estimé qu'elles représentent entre 5 et 10 % des mots d'une conversation de ce type [Shriberg, 2001].

Les phénomènes d'hésitation correspondent généralement à un travail de formulation de la part du locuteur, dans la mesure où le discours est composé au fur et à mesure de sa production. Ils peuvent cependant avoir d'autres rôles tels que la manifestation d'un doute, l'envoi d'un signal au récepteur pour garder la parole ou au contraire la lui céder, le marquage de frontières dans le discours, ou bien encore l'expression d'un stress ou d'une décontraction.

Comme les marques d'hésitation introduisent une rupture dans la continuité du discours, ou plus exactement dans le déroulement syntagmatique, elles sont également appelées *disfluences*. Il existe bien d'autres termes utilisés dans la littérature pour les nommer, certains ne désignant pas exactement les mêmes phénomènes : *turbulences*, *faux départs*, *lapsus*, *inattendus structurels*, *spontanéités*, *modes de production de la langue parlée*, *marques de réparation*, *marques du travail de formulation*, *extragrammaticalités*, *etc.* Une telle profusion de termes témoigne de la discordance entre certains linguistes sur ce qui relève ou non de l'hésitation à l'oral. Toutefois, depuis quelques années, plusieurs études se sont intéressées aux phénomènes d'hésitation et une nomenclature commence à se dégager.

Ces phénomènes peuvent être de natures différentes. Ils incluent par exemple les *pauses silencieuses*, mais pas toutes, certaines jouant d'autres rôles que l'hésitation comme la hiérarchisation et la structuration des constituants, ou encore la mise en valeur stylistique de certains syntagmes. Les *pauses remplies*, correspondant au « *um* » ou au « *uh* » en anglais et au « *euh* » en français, ou encore les *allongements vocaliques* en fin de mots sont des marques d'hésitation très employées [Henry, 2002 ; Candea, 2000]. On peut aussi citer [Shriberg, 2001] :

- les répétitions : « *tous les - les jours*[3] »,
- les suppressions : « *c'est - il est arrivé lundi* »,
- les substitutions : « *tous les jours - toutes les semaines* »,
- les insertions : « *je suis convaincu - je suis intimement convaincu* »,
- les erreurs d'articulation : « *en jouin - en juin* ».

Ces quatre derniers types de disfluence sont parfois désignés sous le terme d'autocorrections [Henry, 2002].

Il est possible de faire une autre classification des phénomènes d'hésitation, en distinguant [Pallaud et Henry, 2004] :

- les bribes, correspondant à une reprise à partir de syntagmes inachevés : « *il a quand même un - une fibre pédagogique* »,
- les amorces, associées aux mots inachevés : « *c'est pas malho - c'est pas malhonnête* ».

[3] Le « - » représentant une pause silencieuse.

Les différents phénomènes d'hésitation sont rarement produits seuls mais plutôt en combinaison. Il peut ainsi y avoir des répétitions de fragments de mots (« *on le re- re- revendique encore une fois* ») [Henry et Pallaud, 2003]. De même, pour assurer un rôle d'hésitation, les pauses silencieuses sont généralement associées à des allongements vocaliques ou à des pauses remplies [Campione et Véronis, 2004].

Les disfluences viennent perturber la modélisation acoustique des mots ainsi que le calcul des probabilités de séquences de mots par les ML. Les amorces et les contractions de mots sont particulièrement difficiles à modéliser puisque le vocabulaire du système de RAP a une taille limitée et ne les inclut généralement pas. Certaines marques d'hésitation, notamment les pauses remplies et les répétitions, sont assez facilement détectables par des méthodes automatiques et peuvent donc être omises dans le calcul des probabilités ou supprimées en vue d'une analyse ultérieure. Toutefois, les disfluences ne peuvent être vues seulement comme du bruit et peuvent informer sur la structure syntaxique du groupe de souffle. Les pauses remplies par exemple aident à la prédiction du mot suivant [Stolcke et Shriberg, 1996] et sont donc souvent considérées comme des mots à part entière par les systèmes de RAP.

Cette section a présenté les principales particularités de l'oral, souvent liées à un contexte d'utilisation moins formel que celui de l'écrit. Les différences observées au niveau du vocabulaire employé et de la syntaxe sont particulièrement manifestes pour l'oral spontané, marqué par une présence importante de disfluences. Ces caractéristiques de l'oral perturbent l'utilisation de méthodes traditionnellement employées en TAL sur des textes écrits, ce qui complique l'acquisition de connaissances linguistiques à partir des premières hypothèses de transcription. La section suivante présente les types d'informations qui peuvent être employés pour réduire le WER, en précisant les méthodes envisagées pour les extraire et les exploiter, et leur influence sur la qualité de la transcription produite.

2.2 Quelles connaissances linguistiques ?

Si l'on se réfère aux modules faisant partie intégrante d'un système de RAP (*cf.* section 1.1), deux semblent pouvoir bénéficier de connaissances linguistiques supplémentaires. Le MA tout d'abord se base sur des indices acoustiques pour reconnaître des phones et gagnerait donc à utiliser des connaissances additionnelles sur le plan de la phonétique et de la phonologie[4]. La plupart des systèmes de RAP intègrent dorénavant les contextes de prononciation dans la reconnaissance des sons ; ce ne sont plus ainsi des phones qui sont à identifier mais des triphones, représentant un phone dans des contextes particuliers gauche et droit. Des travaux ont également envisagé la reconnaissance de traits phonétiques spécifiques pour guider ensuite le décodage de la parole [Liu, 1995 ; Juneja, 2004].

[4]La phonétique s'intéresse à la production physique et à la perception des phones, tandis que la phonologie étudie comment sont agencés les phonèmes d'un langage pour former des énoncés.

Le ML, quant à lui, est par définition la partie du système de RAP qui introduit des connaissances sur la langue. Il peut en outre intégrer des informations plus diversifiées que le MA. Par conséquent, c'est principalement lui qui a vocation à inclure des connaissances linguistiques supplémentaires et nous n'envisageons dans la suite de ce mémoire que les informations prises en compte à son niveau.

Cette section fait un tour d'horizon des types d'informations linguistiques qui ont été pris en compte en RAP au sein du ML. Nous y évoquons succinctement les méthodes de TAL permettant de les acquérir et y exposons les performances obtenues en terme de WER. La présentation est structurée selon une typologie à cinq niveaux généralement reconnue en linguistique :

- la *phonologie* et la *phonétique* qui étudient les sons,
- la *morphologie* qui étudie la structure des mots,
- la *syntaxe* qui étudie la structure des syntagmes et des phrases,
- la *sémantique* qui étudie les sens des mots, des locutions, des phrases ou des textes,
- la *pragmatique* qui étudie la relation entre le langage et son contexte d'utilisation.

Il est à noter que les techniques d'intégration de nouvelles informations sur la langue se heurtent à plusieurs difficultés ; elles doivent être robustes aux distorsions de l'oral et pouvoir s'intégrer dans un ML statistique pour sélectionner les hypothèses les plus probables. Ce dernier point complique notamment la prise en compte de connaissances symboliques au cours de la transcription [Antoine et Genthial, 1999].

2.2.1 Phonologie et phonétique

Les ensembles de phonèmes et de phones d'une langue sont finis et il est possible d'établir manuellement des règles quant à la succession de leurs éléments. En français, l'emploi consécutif des deux phones [d] et [s] est par exemple illicite. De même, les règles d'élision imposent que certains mots terminés par une voyelle ne peuvent être suivis par une voyelle, comme dans l'exemple « *le arbre* ».

Peu d'informations de ce type ont été intégrées dans un ML. Une première étude que nous avons rencontrée prend en compte des règles phonologiques contextuelles lors du décodage en pénalisant les suites incorrectes de prononciations de mots [Gravier *et al.*, 2001]. Cette connaissance permet de réduire la taille de l'espace de recherche en écartant des prononciations incorrectes, comme par exemple l'erreur de liaison [le aʁbʁ][5] commise sur la suite de mots « *les arbres* ». Une des limites à l'emploi de ce type de connaissance repose sur le manque de ressources disponibles annotées phonétiquement. Une seconde étude porte sur la détermination des événements impossibles [Langlois *et al.*, 2003]. Cette méthode consiste à affiner le calcul des probabilités des modèles N-grammes, en excluant du langage $\mathcal{L}$ toutes les séquences de mots illicites. En éliminant les bigrammes ne respectant pas les règles d'élision, seul 0,1 % de l'ensemble des bigrammes s'est retrouvé écarté, ce qui n'apporte donc pas de gain significatif en terme de qualité de la transcription.

[5] En lieu et place de [lɛz aʁbʁ].

2.2.2 Morphologie

La structuration des mots constitue une source d'information intéressante pour des langues morphologiquement très riches comme le turc, le russe ou l'arabe ou même pour des langues à haut taux de flexion comme le français. Dans le cas des langues agglutinantes notamment, pour avoir une couverture lexicale similaire à celle obtenue pour l'anglais, il est nécessaire d'envisager un nombre considérable de mots ; le fait de décomposer les mots en plusieurs constituants élémentaires permet de réduire le nombre d'événements à envisager lors du calcul des probabilités. Dans le cas des langues flexionnelles, l'analyse morphologique permet de rassembler dans une même classe des mots qui sont très liés sémantiquement, comme « *mangeait* » et « *mangerons* » en français.

L'acquisition des informations morphologiques à partir de la transcription pose peu de problèmes particuliers par rapport aux documents écrits puisque les mots reconnus appartiennent à un vocabulaire prédéfini et ont ainsi une orthographe correcte. Si les connaissances se limitent aux lemmes, des lemmatiseurs peuvent être employés [Namer, 2000]. Une variante peut consister à employer des *stemmers* [Porter, 1980 ; Savoy, 1999], qui à la différence des lemmatiseurs ne retournent pas des mots sous une flexion précise, *e.g.* « *imaginer* » pour « *imagina* », mais des formes canoniques, *e.g.* « *imagin* », qui peuvent être des mots tronqués. Cette dissemblance est liée au mode de fonctionnement des *stemmers* qui se limitent à des règles de réécriture. Une autre différence réside dans le fait que les lemmatiseurs limitent l'équivalence entre deux mots à la morphologie flexionnelle, *e.g.* entre « *fabriquera* » et « *fabriquait* », tandis que les *stemmers* examinent également la morphologie dérivationnelle en ramenant par exemple « *fabriquer* » et « *fabrication* » à une même forme. Dans le cas où l'on souhaite séparer un mot en différents constituants et non plus le ramener à une forme simple, d'autres techniques permettent de découper un mot en morphes[6] ou en morphèmes[7] [Allen, 1976 ; Çarki *et al.*, 2000] ; elles sont principalement utilisées pour les langues agglutinantes. La reconnaissance de ces unités morphologiques pouvant être difficile à faire automatiquement, les mots sont parfois simplement découpés en des segments qui correspondent à des définitions linguistiques beaucoup plus vagues que les notions de morphe et de morphème [Whittaker, 2000].

Les connaissances sur la morphologie peuvent être intégrées au système de RAP au moyen des modèles N-classes. Des ML ont ainsi été construits en réunissant tous les mots associés à un même lemme dans une même classe [El-Bèze et Derouault, 1990 ; Maltese et Mancini, 1992]. Leur comportement n'a pas été toutefois testé dans des systèmes de RAP. Une alternative aux modèles N-classes est le *ML factorisé*. Ce ML décompose chaque mot w_i en k caractéristiques, aussi appelées facteurs [Kirchhoff *et al.*, 2007]. Elles représentent des informations morphologiques, syntaxiques ou sémantiques sur le mot, en plus du mot lui-même. Ces ML factorisés ont notamment été appliqués

[6] Segment d'un mot qui est un signe élémentaire, *i.e.*, qui ne peut être décomposé en plusieurs autres signes de la langue. Ainsi, le mot « *chiens* » est constitué des morphes « *chien* » et « *s* ».

[7] Regroupement de morphes ayant le même signifié. Par exemple, les deux morphes « *majeur* » et « *major* » sont regroupés sous le même morphème « {MAJEUR} ». La distinction entre morphe et morphème est du même ordre que celle faite entre phone et phonème.

à l'arabe, avec comme facteurs le radical[8], la racine[9] et la classe morphologique. Leur utilisation au sein d'un système de RAP pour transcrire des émissions d'actualité a permis de réduire le taux d'erreur sur les mots de 57,6 % à 56,1 % [Vergyri *et al.*, 2004]. Elle se heurte toutefois à deux difficultés. La première concerne l'application du repli dans ce cadre. Un mot étant découpé en plusieurs facteurs, il n'y a plus d'ordres prédéfinis entre les facteurs, ce qui ne se produit pas dans le cadre des ML classiques où deux mots ne peuvent pas être émis simultanément. Il devient ainsi nécessaire, avec les ML factorisés, de développer des méthodes pour déterminer les éléments à supprimer en premiers dans l'historique pris en compte par les probabilités conditionnelles. La seconde difficulté est relative à l'intégration de ces modèles dans le processus de transcription. Dans la mesure où les décodeurs opèrent habituellement sur des mots et non sur des facteurs, les ML factorisés demandent des modifications importantes du système de RAP et entraînent un accroissement considérable des calculs quand les mots doivent être étendus en plusieurs versions de facteurs en cas d'ambiguïté. Pour ces raisons, ces modèles sont utilisés en post-traitement d'un système de RAP et leur emploi se limite aux langues morphologiquement très riches.

2.2.3 Syntaxe

L'information syntaxique peut être utilisée sous plusieurs formes : les ML peuvent ainsi tenir compte des classes grammaticales attribuées aux mots, d'une nouvelle segmentation du groupe de souffle en réunissant plusieurs mots au sein d'un même constituant, ou encore d'une analyse syntaxique du groupe de souffle. Nous présentons successivement un état de l'art des tentatives faites pour intégrer ces divers types de connaissances.

2.2.3.1 Parties du discours

Comme déjà dit en introduction générale, une *partie du discours* ou *POS* (pour *part of speech*) est une catégorie grammaticale qui est attribuée à un mot, voire à un groupe de mots dans le cas de locution, en fonction de son contexte d'utilisation. Traditionnellement, neuf classes sont distinguées dans les grammaires des langues naturelles : les adjectifs, les adverbes, les articles, les conjonctions, les interjections, les noms, les prépositions, les pronoms et les verbes. Chacune d'entre elles est en outre subdivisée. Ainsi parmi les conjonctions retrouve-t-on les conjonctions de coordination et de subordination ; de la même manière, on distingue généralement les noms communs des noms propres. De plus, les classes grammaticales associées aux mots peuvent être accompagnées d'informations morphologiques sur le genre, le nombre, le personne, le temps, le mode ou encore le cas employé ; les classes sont alors plutôt désignées comme étant

[8]Support morphologique d'un mot. C'est la partie qui contient le sens d'un mot, après avoir supprimé tout ce qui relevait de la flexion. Par exemple, les radicaux respectifs de « *déstabiliser* » et « *nationaliser* » sont « *déstabilis-* » et « *nationalis-* ».

[9]Constituant d'un mot qui porte la partie principale de son sens. À la différence du radical, la racine ne peut pas être décomposée en d'autres éléments porteurs de sens ou morphologiquement simples. Ainsi, les racines respectives de « *déstabiliser* » et « *nationaliser* » sont « *stabil-* » et « *nation* ».

morpho-syntaxiques bien que le terme de POS est également souvent utilisé dans ce cadre.

Les POS sont affectées aux mots selon le rôle syntaxique qu'ils occupent dans la phrase. Il existe de ce point de vue deux ensembles bien disjoints de classes grammaticales. On trouve d'une part les mots *lexicaux*, qui portent une charge lexicale pleine et qui appartiennent à des classes dites ouvertes, car ils sont en grand nombre dans une langue et sont en création perpétuelle ; il s'agit des noms, des adjectifs, des verbes et des adverbes. D'autre part figurent les mots *grammaticaux* ou *mots-outils* qui participent à la structuration de la phrase, sont décrits plutôt par leur rôle syntaxique que par leur sémantique et sont en nombre limité dans une langue, ce qui leur vaut d'appartenir à des classes dites fermées. Dans cette catégorie, on peut classer les déterminants, les prépositions, les conjonctions, les pronoms et les verbes auxiliaires. Les interjections forment quant à elles une classe à part, en n'ayant pas de connexions grammaticales avec le restant de la phrase.

La granularité et la hiérarchie dans la taxonomie des POS est sujette à discussion. Suivant l'application visée, on subdivisera plus ou moins chacune des neuf classes. En outre, ces grandes catégories engendrent quelques difficultés puisque par exemple les adjectifs non qualificatifs comme « *mon* » sont plus proches des articles que des adjectifs qualificatifs.

Les POS peuvent être attribuées automatiquement au moyen des étiqueteurs morpho-syntaxiques ; le chapitre 4 est consacré à l'application de ces outils sur l'oral et nous ne nous étendrons pas davantage ici sur leur fonctionnement. Elles peuvent être utilisées comme informations supplémentaires au sein des ML en remarquant qu'il existe des successions de POS qui se produisent de manière rarissime. En français, un déterminant ne peut ainsi jamais suivre un nom commun. Une autre justification de leur emploi est qu'en rassemblant plusieurs mots partageant des propriétés syntaxiques communes au sein d'une même classe, le nombre d'événements possibles à envisager diminue par rapport aux mots considérés isolément.

Les POS sont généralement intégrées pour ces raisons aux systèmes de RAP au moyen de modèles N-classes, où chaque POS correspond à une classe [Maltese et Mancini, 1992]. Les modèles N-classes sont moins performants que les modèles N-grammes, en considérant des historiques de même longueur. On observe en revanche dans certains cas une amélioration de la perplexité par rapport aux modèles N-grammes quand on interpole linéairement (*cf.* section 2.3.2) des modèles N-classes avec des modèles N-grammes. Cette baisse de la perplexité reste cependant peu importante même quand on dispose de suffisamment de données pour apprendre les paramètres du ML [Weintraub *et al.*, 1996] ; diverses améliorations des modèles N-classes ont donc été envisagées.

Dans une première solution, le mode d'intégration dans le système de RAP des modèles N-classes utilisant les POS est révisé. Au lieu d'être employé sur le graphe de mots produit aux cours des passes précédentes, le ML est appliqué de manière plus sélective sur un graphe d'homophones [Béchet *et al.*, 1999]. Chaque mot de la meilleure

hypothèse est alors remplacé par tous ses homophones[10] possibles, *i.e.*, par tous les mots du langage qui ne sont pas distinguables au moyen d'indices acoustiques. Ce mode de représentation est particulièrement adapté pour corriger des fautes d'accord en genre et en nombre, notamment en français où les formes d'un même mot au singulier et au pluriel sont souvent homophones. L'application d'un modèle N-classes adoptant cette méthode a permis de réduire le taux d'erreur de 10,7 % à 10,5 % sur la transcription d'émissions d'actualité en français [Gauvain *et al.*, 2005].

Une deuxième solution consiste à reconsidérer le mode de calcul des probabilités. En modifiant l'approximation faite sur la probabilité lexicale dans l'équation (1.18), page 22 par :

$$P(w_i|w_1^{i-1}, c_1^i) \approx P(w_i|c_{i-N+1}^i) \ , \tag{2.1}$$

une légère amélioration de l'entropie croisée a été observée [Goodman, 2001]. On peut même aller jusqu'à supprimer les approximations faites à la fois sur les probabilités lexicale et contextuelle. Une étude [Heeman, 1999] propose ainsi de redéfinir l'objectif d'un système de RAP (*cf.* section 1.1), de manière à ce que les POS C associées aux mots W à reconnaître soient considérées comme partie intégrante de la sortie de la transcription et non plus comme des objets intermédiaires. La fonction objectif de la RAP revient alors à estimer :

$$\hat{W}, \hat{C} = \arg\max_{W,C} P(W, C|A) \ . \tag{2.2}$$

En éliminant les approximations, le nombre d'événements à examiner pour évaluer les probabilités augmente considérablement ; une méthode basée sur des arbres de décision a donc été élaborée. Cette technique a donné des résultats satisfaisants pour transcrire des dialogues portant sur des sujets précis puisque le ML triclasses modifié a conduit à une réduction du taux d'erreur de 26,0 % à 24,9 % par rapport à des ML trigrammes, tandis que le ML triclasses conventionnel faisait quant à lui augmenter le taux d'erreur.

Une troisième solution pour améliorer les modèles N-classes est de modifier la construction des classes. Au lieu d'avoir une classe par POS, une possibilité est de regrouper au sein d'une même classe l'ensemble des mots ayant les mêmes POS possibles avec le même ordre de vraisemblance. Ceci supprime l'ambiguïté des classes que l'on peut associer aux mots. Bien que l'attribution d'une POS à un mot soit une technique relativement maîtrisée (*cf.* section 4.2), il demeure toujours en effet des erreurs qui peuvent venir perturber le calcul des probabilités. Un ML triclasses utilisant ce type de classes, combiné avec un ML trigrammes, a permis de réduire le taux d'erreur sur les mots de 12,3 % à 11,6 % ou de 9,8 % à 9,4 % selon le MA utilisé, pour traiter de la parole lue [Samuelsson et Reichl, 1999].

On peut également envisager un système à nombre très restreint de classes, en poussant à l'extrême la propriété principale des modèles N-classes, qui est de réduire le total des événements à envisager pour le calcul des probabilités. Généralement, une vingtaine à une centaine de POS différentes sont considérées ; toutefois, seuls deux

[10] L'homophonie désigne le rapport entre deux mots possédant la même prononciation mais des graphies différentes, *e.g.* entre « *mère* » et « *mer* ».

types de POS sont parfois distingués : les mots lexicaux et les mots grammaticaux. Cette distinction est faite avec l'idée que la séquence des mots grammaticaux reflète les contraintes syntaxiques du groupe de souffle, alors que la séquence de mots lexicaux est contrôlée par des relations sémantiques entre les mots. Les classes ouvertes et fermées sont généralement utilisées en adaptant les modèles N-classes [Isotani et Matsunaga, 1994 ; Geutner, 1996]. Dans le cas où on considère un historique de longueur deux, seuls les derniers mots lexical et grammatical rencontrés sont considérés. Si o_{i-1} est le dernier mot de classe ouverte et f_{i-1} le dernier mot de classe fermée dans w_1^{i-1}, le calcul des probabilités se fait comme suit :

$$P(w_i|w_1^{i-1}) \approx \begin{cases} P(w_i|w_{i-1}, o_{i-1}) & \text{si } w_{i-1} \text{ est un mot de classe fermée} \\ P(w_i|w_{i-1}, f_{i-1}) & \text{si } w_{i-1} \text{ est un mot de classe ouverte} \end{cases} \quad (2.3)$$

L'interpolation linéaire d'un tel modèle avec un ML trigrammes a permis de réduire le taux d'erreur de 29,4 % à 29,0 % pour transcrire un corpus de parole spontané en allemand [Geutner, 1996]. La prise en compte de ces deux types de classes peut être faite différemment en utilisant un modèle N-grammes classique pour les classes ouvertes et un modèle spécialement conçu pour les mots de classes fermées [Peng et Schuurmans, 2001]. Le modèle spécifique prédit les mots grammaticaux à partir des $M-1$ mots grammaticaux précédents. La conception de ce modèle est justifiée par le fait qu'en anglais par exemple, les mots des classes fermées représentent 30 % du langage écrit et sont en moyenne distants de 1,9 mots. L'utilisation de ce modèle a permis une légère amélioration de la perplexité par rapport à un ML trigrammes.

Outre les modèles N-classes, la connaissance sur les POS peut être introduite grâce aux modèles de cache. Un cache, limité à 200 mots, peut être construit pour chaque POS [Kuhn et De Mori, 1990]. Le cache d'une POS donnée contiendra alors les derniers mots rencontrés étiquetés par cette POS. Ces modèles sont conçus avec l'idée que chaque POS a une répartition particulière d'occurrences. Les mots lexicaux ont ainsi tendance à apparaître par vagues, au gré des sujets traités, tandis que les mots grammaticaux sont répartis plus uniformément. Ce type de modèle possède des propriétés intéressantes puisque la combinaison d'un modèle de cache avec un ML trigrammes a conduit à une réduction de perplexité d'un facteur supérieur à trois par rapport à un modèle trigrammes [Kuhn et De Mori, 1990]. Un modèle assez similaire a été également conçu pour discriminer les formes singulier et pluriel homophones [Béchet *et al.*, 1999].

2.2.3.2 Multimots

Une autre utilisation de connaissances syntaxiques consiste à regrouper plusieurs mots au sein d'unités d'ordre supérieur, comme des locutions ou des syntagmes, et à les ajouter au vocabulaire du système de RAP. Ces groupes de mots, que nous désignons par la suite par le terme *multimots*, peuvent être de natures très diverses. Il peut s'agir de mots qui cooccurrent fréquemment dans un corpus, tels que « *demain matin* » ou encore « *millions de dollars* », de mots composés comme « *New York* » ou « *vice président* », ou d'entités nommées concernant des dates ou des noms de personne. Dans le cas

d'applications ciblées, ce peut être des expressions propres à un domaine telles que « *vous êtes la bienvenue* » ou « *pouvez-vous s'il vous plaît me mettre en contact avec* ».

Les multimots sont généralement sélectionnés par des méthodes automatiques parmi l'ensemble des combinaisons possibles de mots du vocabulaire du système de RAP. Il existe deux approches principales pour les obtenir : les méthodes purement statistiques et les systèmes basés sur des indices structurels, même si cette distinction est parfois un peu arbitraire, certaines techniques combinant les deux approches [Béchet *et al.*, 1999]. Parmi les méthodes purement statistiques figurent les multigrammes, déjà évoqués en section 1.2.2 ; on peut également mentionner des techniques sélectionnant les multimots suivant un critère donné, tel que les fréquences de cooccurrences ou l'évaluation de la perplexité [Suhm et Waibel, 1994 ; Ries *et al.*, 1996 ; Kuo et Reichl, 1999 ; Zitouni *et al.*, 2003]. Parmi les méthodes basées sur des indices structurels, on peut citer celles utilisant des automates à états finis [Nasr *et al.*, 1999 ; Béchet *et al.*, 1999] ou des grammaires non contextuelles probabilistes [Gillett et Ward, 1998 ; Wang *et al.*, 2000 ; Mou *et al.*, 2002 ; Seneff *et al.*, 2003]. L'ajout de multimots au vocabulaire ayant l'inconvénient d'augmenter le nombre d'événements possibles, des modèles N-classes sont souvent envisagés. Les classes peuvent indifféremment contenir des mots et des multimots [Ries *et al.*, 1996 ; Deligne et Sakisaga, 1998 ; Zitouni *et al.*, 2003] ou bien n'être construites que pour les multimots [Nasr *et al.*, 1999 ; Gillett et Ward, 1998 ; Wang *et al.*, 2000 ; Mou *et al.*, 2002 ; Seneff *et al.*, 2003].

L'intérêt principal des multimots par rapport aux mots est d'introduire des connaissances phonétiques précises sur la prononciation de séquences courantes, notamment au niveau des liaisons entre les mots. Un autre avantage est qu'ils autorisent la prise en compte de phénomènes entre mots distants, tels que les accords en genre et en nombre, sans avoir à augmenter la taille de l'historique des ML [Béchet *et al.*, 1999]. Ils permettent aussi d'introduire des connaissances spécifiques à un domaine en indiquant des expressions à reconnaître [Kuo et Reichl, 1999] ou encore des règles pour identifier certaines entités nommées [Mou *et al.*, 2002]. Il semble d'ailleurs que l'intérêt des multimots pour la RAP dépende directement du domaine étudié puisque si dans des applications ciblées telles que le routage d'appels téléphoniques, les constructions stéréotypées telles que « *can you please get me* » sont très naturellement modélisées, cela se révèle plus délicat dans le cadre de la parole lue, où les constructions et le vocabulaire sont riches et variés. Des expériences menées sur la transcription de la lecture d'un corpus aussi général que celui du *Monde* montre toutefois une amélioration relative du WER de 14 % en recourant à des multimots, par rapport à un système limité à des ML bigrammes et trigrammes classiques [Zitouni *et al.*, 2003]. L'usage des multimots est en tout cas moins systématique en RAP que dans le domaine de la traduction statistique où certains systèmes sont entièrement basés sur le repérage de ces groupes de mots [Koehn *et al.*, 2003].

2.2.3.3 Analyse syntaxique

Le principal avantage de l'analyse syntaxique est de prendre en compte les dépendances syntaxiques entre les constituants d'un même groupe de souffle, et ce, même si

ces constituants se trouvent à des positions assez éloignées. Par exemple, pour calculer le score de la séquence de mots « *les chats de la voisine du dessus sont noirs* », la construction d'un arbre d'analyse syntaxique semble appropriée pour établir la relation entre « *chats* », qui est la tête du groupe nominal « *les chats de la voisine du dessus* », et le verbe « *sont* ».

Les méthodes d'analyse syntaxique sont toutefois confrontées à une difficulté importante, à savoir la conception de grammaires suffisamment robustes. Il est déjà difficile d'élaborer pour l'écrit des analyseurs syntaxiques ayant une large couverture, même si certaines grammaires lexicalisées parviennent à une précision[11] et un rappel[12] proches de 90 % pour analyser une partie du corpus du *Wall Street Journal* [Charniak, 2000]. Les difficultés intrinsèques à l'oral (*cf.* section 2.1), notamment le manque de ponctuation, la présence de disfluences et l'absence éventuelle de majuscules dans la transcription, ajoutées aux erreurs de reconnaissance des systèmes de RAP, compliquent encore la réalisation d'analyseurs. La campagne d'évaluation EASY, menée sur différents types de documents en langue française, illustre ces difficultés en montrant une importante dégradation de l'analyse syntaxique dès lors que le texte à traiter contient de nombreuses disfluences [Paroubek *et al.*, 2006].

L'utilisation de l'analyse syntaxique par les systèmes de RAP a, pour ces raisons, longtemps été confinée à des applications homme-machine où les tournures de phrase et le vocabulaire étaient très spécifiques [Seneff, 1992]. Les méthodes d'analyse partielle, qui ne nécessitent pas de construire un arbre syntaxique décrivant la structure détaillée du groupe de souffle entier, sont particulièrement adaptées pour concevoir des solutions robustes. Un ML a ainsi été défini en segmentant les suites de mots à analyser en constituants non récursifs, que l'on appelle *chunks* [Zechner et Waibel, 1998]. Une difficulté supplémentaire à l'utilisation de méthodes d'analyse syntaxique est leur lenteur ; dans le cas notamment de la construction d'arbre d'analyse syntaxique, plusieurs peuvent être construits en parallèle pour trouver celui qui correspond le mieux au groupe de souffle courant et cela ralentit le décodage.

Malgré toutes ces difficultés, plusieurs tentatives ont été faites. Les *grammaires non contextuelles*, notamment leurs versions probabilistes, ont été le premier type d'analyse envisagé dans les systèmes de RAP [Stolcke et Segal, 1994 ; Jurafsky *et al.*, 1995 ; Seneff, 1992 ; Seneff *et al.*, 1995 ; Linares *et al.*, 2004]. Le mode d'attribution des probabilités pour ces grammaires est souvent jugé limité puisqu'il ne dépend que du non-terminal de la partie gauche de chacune des règles et non des mots eux-mêmes ; d'autres formes d'analyse sont souvent préférées. Les *grammaires lexicalisées* [Jurafsky et Martin, 2008] étendent ainsi les grammaires non contextuelles en choisissant pour chaque constituant détecté un mot jouant le rôle de tête. Le calcul des probabilités est alors conditionné par la tête [Chelba et Jelinek, 2000] et parfois par d'autres non-terminaux rencontrés auparavant dans l'analyse [Roark, 2001 ; Charniak, 2001]. Les *grammaires de liens* [Lafferty *et al.*, 1992 ; Berger et Printz, 1998] effectuent quant à elles

[11] Définie par $\frac{\text{nb de constituants communs à GOLD et à TEST}}{\text{nb de constituants dans TEST}}$, où GOLD et TEST sont les arbres d'analyse obtenus respectivement manuellement et automatiquement.

[12] Défini par $\frac{\text{nb de constituants communs à GOLD et à TEST}}{\text{nb de constituants dans GOLD}}$.

une analyse syntaxique en déterminant les liens entre les mots, tandis que les *grammaires de dépendance par contraintes* [Harper et Helzerman, 1995 ; Wang et Harper, 2002] expriment les dépendances existant au sein d'une phrase par des contraintes syntaxiques et sémantiques. Si les règles de toutes ces grammaires sont généralement déterminées *a priori*, l'évaluation des probabilités du ML se fait automatiquement soit à partir de corpus annotés syntaxiquement tels que le Penn Treebank, soit à partir de corpus analysés par des méthodes automatiques.

La manière dont sont prises en compte les techniques d'analyse syntaxique dépend de la grammaire utilisée. Lorsque ce sont des versions non probabilistes qui sont employées, une solution peut consister à examiner la proportion de mots du groupe de souffle couverts par l'analyse ; plus ce nombre sera grand, meilleure sera considérée l'hypothèse [Zechner et Waibel, 1998]. Quand ce sont des grammaires probabilistes que l'on envisage, les probabilités contextuelles du ML sont évaluées au moyen de :

$$P(w_i|w_1^{i-1}) = \frac{P(w_1^i)}{P(w_1^{i-1})} \quad . \tag{2.4}$$

En ce qui concerne les grammaires non contextuelles probabilistes, il existe des algorithmes rapides [Jelinek et Lafferty, 1991 ; Stolcke, 1995] permettant de calculer avec exactitude les probabilités $P(w_1^i) = P(S \stackrel{*}{\rightarrow} w_1 w_2 \dots w_i \dots)$. Une autre technique, employée aussi bien par les grammaires non contextuelles que par les lexicalisées, repose sur les probabilités associées aux arbres de dérivation [Linares *et al.*, 2004 ; Roark, 2001 ; Charniak, 2001]. Si $D_{w_1^i}$ représente l'ensemble des dérivations associées à w_1^i, on peut déterminer $P(w_1^i)$ par :

$$P(w_1^i) = \sum_{d \in D_{w_1^i}} P(d) \tag{2.5}$$

où $P(d)$ est obtenue en faisant le produit des probabilités de toutes les règles utilisées dans la dérivation d. En pratique, l'ensemble $D_{w_1^i}$ peut être très grand, ce qui conduit à conserver uniquement les dérivations les plus probables.

Les grammaires ont été principalement intégrées dans les systèmes de RAP pour des applications ciblées ; les non contextuelles probabilistes ont ainsi permis de diminuer le taux d'erreur dans les domaines de réservations pour le trafic aérien (de 34,6 % à 29,6 % par rapport à un modèle bigrammes) [Jurafsky *et al.*, 1995] et de la restauration (de 6,9 % à 6,7 % par rapport à un modèle quadrigrammes) [Seneff *et al.*, 1995]. Les progrès récents de l'analyse syntaxique, obtenus en prenant en compte un contexte de plus en plus précis lors de l'attribution des probabilités à chacune des règles, ont autorisé son utilisation sur de la parole lue ; des grammaires lexicalisées ont ainsi conduit à une baisse de perplexité par rapport aux modèles trigrammes sur un corpus du *Wall Street Journal* [Chelba et Jelinek, 2000 ; Roark, 2001 ; Charniak, 2001]. Les modèles conçus avec ces grammaires sont cependant trop coûteux en calculs pour être utilisés au sein de systèmes de RAP dans des applications à grande échelle, même si l'un de ces modèles a pu être intégré dans un ML en réordonnant des graphes de mots [Hall et Johnson, 2004]. L'intégration des grammaires de dépendance par contraintes semble quant à elle

plus aisée puisqu'elle a permis d'obtenir des ML de complexité raisonnable, et ce, en constatant une diminution du taux d'erreur pour transcrire des émissions d'actualité (de 14,7 % à 14,3 % par rapport à un modèle trigrammes) [Wang *et al.*, 2003a] mais aussi des conversations téléphoniques (avec une baisse relative de 6,2 % par rapport à des ML 4-grammes et 4-classes interpolés) [Wang *et al.*, 2004].

2.2.4 Sémantique

L'introduction de connaissances sémantiques vise à favoriser les hypothèses qui possèdent plusieurs mots proches au niveau de leur sens, en supposant que les groupes de souffle à reconnaître ont une certaine cohérence sémantique. On attribuera ainsi un meilleur score à des hypothèses de décodage contenant les mots « *action* » « *obligation* » et « *bourse* » qu'à une hypothèse n'ayant pas de champ sémantique bien défini.

Les relations sémantiques existant entre les mots sont établies soit à partir de ressources telles que WordNet [Fellbaum, 1998], soit automatiquement à partir de corpus [Rossignol, 2005]. La désambiguïsation automatique du sens d'un mot selon son contexte d'utilisation reste encore aujourd'hui un problème difficile [Ide et Véronis, 1998] ; elle l'est encore davantage dans le cas de documents transcrits automatiquement puisque le contexte de chaque mot devient bruité par les erreurs de reconnaissance. Dans [Demetriou *et al.*, 1997], le problème est par exemple contourné en fusionnant les descriptions sémantiques de tous les sens associés à un mot ambigu.

Les connaissances sémantiques ont été globalement peu envisagées pour améliorer la transcription. Une première possibilité pour introduire ce type d'informations repose sur l'utilisation de connaissances *a priori*. Un dictionnaire, contenant les domaines d'emplois et les définitions des sens de 36 000 lemmes en anglais, a ainsi permis d'établir des associations sémantiques entre paires de mots et à sélectionner les hypothèses qui sont les plus homogènes du point de vue sémantique [Demetriou *et al.*, 1997]. Des expériences ont montré que le modèle ainsi construit était informatif lorsque le groupe de souffle à analyser était long mais il n'a pas été comparé à des ML N-grammes.

Une autre technique étudiée pour introduire des connaissances sémantiques est l'*analyse sémantique latente* [Bellegarda, 1998]. Le principe de cette méthode, utilisée en recherche d'information, repose sur le calcul d'une matrice de cooccurrences mot-document à partir d'une collection de documents. Cette matrice contient les fréquences avec lesquelles chaque mot w_i apparaît dans un document d_j, *i.e.*, en quelque sorte le lien sémantique que w_i entretient avec d_j. En projetant la matrice dans un espace de dimension plus faible, l'analyse sémantique latente permet de mettre au jour des affinités entre les mots, ce qui permet de définir sur ce principe un nouveau type de ML en déterminant la probabilité conditionnelle $P(w_i|w_1^{i-1})$ à partir de la proximité « sémantique » existant entre w_i et w_1^{i-1}. Une propriété importante de cette technique est qu'elle permet en outre de limiter les problèmes d'éparsité de données en recourant à des techniques de réduction de la dimension. L'analyse sémantique latente est une technique intéressante puisque ce type de ML a permis une baisse relative du taux d'erreur de 16 % par rapport à des ML trigrammes pour transcrire de la parole lue en anglais [Bellegarda, 2000].

2.2.5 Pragmatique

Un système de RAP doit faire face à la grande variabilité du langage (*cf.* section 1.4.1). Selon le locuteur, les sujets abordés, ou encore le cadre de la discussion, on pourra ainsi observer des changements importants du point de vue du vocabulaire, du niveau de langue utilisé, des tournures employées... L'introduction de la pragmatique dans la RAP recouvre un ensemble de techniques qui visent à adapter le système de transcription à ces différences. Deux catégories de méthodes peuvent être distinguées. Si l'on restreint l'adaptation à la modification du système de transcription en fonction du sujet traité, un nouveau ML peut être formé à partir de modèles correspondant chacun à un thème particulier ; une autre possibilité consiste à recourir à un corpus qui se rapporte davantage au signal sonore à traiter que le corpus général.

2.2.5.1 Modèles thématiques

Dans cette approche, des modèles dits thématiques sont établis pour un ensemble fixé de thèmes. Pour décoder un document sonore, les sujets abordés sont dans un premier temps identifiés soit à partir du graphe de mots produit lors d'une passe précédente, soit à partir des mots transcrits avant le mot courant. Dans un second temps, le ou les ML thématiques correspondant aux thèmes détectés sont utilisés, en espérant obtenir une baisse du WER par rapport à un ML général.

La construction des modèles thématiques est réalisée à partir d'un ensemble de documents correspondant chacun à un sous-langage ou à un thème. L'attribution de chaque texte du corpus à une catégorie peut être faite à partir d'une annotation manuelle [Kneser et Steinbiss, 1993 ; Brun, 2003] ou bien en regroupant par des méthodes automatiques les documents similaires [Iyer et Ostendorf, 1999 ; Clarkson et Robinson, 1997 ; Florian et Yarowsky, 1999]. Le nombre d'ensembles distingués dans le corpus d'apprentissage est très variable. Il peut par exemple être inférieur à 10 si les thèmes détectés sont généraux, comme l'économie, l'histoire ou la politique [Brun, 2003 ; Bigi *et al.*, 2000] ou être de plus de 5 000, permettant ainsi une distinction plus fine du sujet abordé [Seymore et Rosenfeld, 1997].

La détection de thème est quant à elle beaucoup plus délicate car elle doit être réalisée à partir de transcriptions automatiques. Les erreurs de substitution (*cf.* section 1.3.2) sont particulièrement problématiques puisqu'elles peuvent remplacer une information utile sur le thème traité par une autre qui est inexacte [Carlson, 1996]. Une autre difficulté liée à la détection est le manque de diversité lexicale des transcriptions. Le vocabulaire d'un système de RAP est en effet de taille finie et est établi en ne retenant que les mots les plus fréquents, qui ne sont pas tous très utiles à la caractérisation des thèmes [Brun, 2003]. La reconnaissance du sujet abordé par le document à transcrire peut se faire à différents niveaux. Elle est souvent réalisée à partir du document entier pour limiter l'influence des erreurs sur les mots reconnus ; elle peut aussi être effectuée au niveau de chaque groupe de souffle [Khudanpur et Wu, 1999 ; Lane *et al.*, 2005] ou d'un ensemble consécutif de groupes de souffles [Chen *et al.*, 2001].

Il existe plusieurs façons d'employer les modèles thématiques au sein d'un système de RAP [Bellegarda, 2004]. Le ML correspondant au thème détecté à partir d'une première hypothèse de transcription peut être ainsi directement utilisé dans la seconde passe du système de RAP [Lane *et al.*, 2005]. Le plus souvent cependant, il est connecté avec un ML général à l'aide d'une interpolation linéaire [Brun, 2003 ; Iyer et Ostendorf, 1999 ; Seymore et Rosenfeld, 1997] ou d'un modèle exponentiel [Khudanpur et Wu, 1999] (*cf.* section 2.3.2). Des méthodes propres à l'adaptation ont également été conçues pour combiner plusieurs modèles, chacun étant spécifique à un thème. Si t_k est un thème présent dans le corpus d'apprentissage, les ML dits à *modèles de mélange* calculent les probabilités de la manière suivante :

$$P(w_i|w_1^{i-1}) = \sum_k \lambda_k(w_1^{i-1})P(w_i|w_1^{i-1}, t_k) \tag{2.6}$$

où les paramètres λ_k sont optimisés de façon à minimiser la perplexité du texte déjà reconnu [Kneser et Steinbiss, 1993 ; Martin *et al.*, 1997 ; Clarkson et Robinson, 1997]. L'intérêt de ce type de modèle est de pouvoir associer plusieurs thèmes au document à transcrire. Une autre solution consiste à exprimer la contribution du thème plus directement :

$$P(w_i|w_1^{i-1}) = \sum_k P(w_k|t_k)P(w_i|w_1^{i-1}, t_k) \tag{2.7}$$

où $P(w_k|t_k)$ représente la probabilité qu'un mot w_k apparaisse dans un document traitant du thème t_k [Gildea et Hofmann, 1999 ; Florian et Yarowsky, 1999].

D'une manière générale, les modèles thématiques ont permis une réduction significative de la perplexité, mais cette réduction s'est traduite par une baisse limitée du taux d'erreur de transcription. Une propriété intéressante est qu'ils semblent apporter des informations complémentaires à celles fournies par une analyse syntaxique. En combinant ces deux sources d'information, il a été ainsi constaté que les gains de chaque source sont presque additifs en ce qui concerne la baisse du taux d'erreur [Wu et Khudanpur, 1999].

2.2.5.2 Corpus d'adaptation

Un autre procédé d'adaptation des ML consiste à modifier le calcul des probabilités en fonction d'un nouveau corpus qui est plus proche du document à transcrire que ne l'est celui qui est utilisé pour l'apprentissage du ML initial. Dans certains cas, ce corpus spécifique peut déjà être disponible mais pas en quantités suffisantes toutefois pour construire un ML complet ayant une bonne couverture du langage. Ainsi, afin de transcrire des débats télévisés, un ML général peut être conçu à partir d'un corpus de grande taille constitué d'émissions d'information puis adapté à partir d'un plus petit corpus constitué en majorité d'interviews [Wang et Stolcke, 2007]. Une autre possibilité de constituer un corpus spécifique est d'extraire uniquement une partie d'un corpus très général, tel celui du journal *Le Monde* [Bigi *et al.*, 2004] ou celui du *Wall Street Journal* [Mahajan *et al.*, 1999 ; Klakow, 2000], en ne conservant que les textes qui se rapportent aux documents à transcrire. D'autres travaux obtiennent des données supplémentaires

aux corpus disponibles en recourant au Web [Berger et Miller, 1998 ; Kemp et Waibel, 1998 ; Nisimura *et al.*, 2001 ; Bulyko *et al.*, 2003 ; Allauzen, 2003 ; Suzuki *et al.*, 2006], ou bien en employant même des transcriptions automatiques produites sur des documents pertinents mais pour lesquels aucune annotation manuelle n'a été réalisée [Gretter et Riccardi, 2001 ; Bacchiani et Roark, 2003].

La modification du système de RAP en fonction d'un corpus d'adaptation peut intervenir à deux niveaux : en remaniant le vocabulaire ou en modifiant le calcul des probabilités des séquences de mots du ML. L'adaptation du vocabulaire est particulièrement pertinente pour transcrire des émissions d'actualité puisque la collecte régulière d'informations sur des sites de dépêches d'agences de presse ou de quotidiens nationaux permet de réduire le taux de mots hors vocabulaire [Kemp et Waibel, 1998 ; Allauzen et Gauvain, 2003]. La conception d'un système de RAP à vocabulaire ouvert a ainsi permis une diminution du WER de 25,5 % à 24,9 % pour la transcription d'émissions d'actualité en français [Allauzen et Gauvain, 2005].

Le calcul des probabilités peut quant à lui être modifié en fonction du corpus d'adaptation selon plusieurs procédés [Bellegarda, 2004]. Une première possibilité consiste à interpoler linéairement un ML appris sur un ensemble de données spécifiques au document traité, avec celui construit sur un corpus général. Les coefficients d'interpolation peuvent alors être déterminés en cherchant à minimiser la perplexité sur le corpus d'adaptation. D'autres techniques consistent à spécialiser un ML général sur le corpus d'adaptation en utilisant un critère de maximum *a posteriori* (MAP) [Federico, 1996 ; Berger et Miller, 1998 ; Chen *et al.*, 2004a] ou un critère de minimum d'information discriminante (MDI pour *Minimum Discrimination Information*) [Federico, 1999 ; Chen *et al.*, 2004a]. L'utilisation de corpus d'adaptation permet une réduction non négligeable du taux d'erreur de la transcription. Il a ainsi été constaté une baisse de 17,1 % à 16,3 % dans la transcription d'émissions d'actualité en anglais. Parmi les techniques précédemment citées, les adaptations au moyen de mélanges dynamiques de modèles et du critère MDI semblent être les plus performantes [Chen *et al.*, 2004a].

À ces méthodes, il convient d'ajouter celles de l'adaptation non supervisée qui consiste à construire un ML à partir d'un corpus sur lequel il n'est pas fait d'hypothèse sur les thèmes qu'il traite. Une méthode actuellement souvent employée est l'allocation de Dirichlet latente [Blei *et al.*, 2003]. Ce modèle, basé sur une représentation sous la forme d'un « sac de mots » des textes étudiés, suppose que chaque document traite plusieurs thèmes et que chaque mot est associé à un des thèmes. Ces thèmes, dont le nombre est fixé expérimentalement, sont des variables latentes de la loi mélange associée à chaque document et sont déterminées lors de l'apprentissage du modèle suivant un indice statistique plutôt que sur des critères linguistiques précis. L'intérêt de ces méthodes d'adaptation non supervisée est de spécifier les distributions de probabilité d'un document de manière plus fine qu'en considérant uniformément le corpus entier. En effet, afin de prédire une nouvelle succession de mots, ces techniques choisissent des thèmes qui correspondent aux extraits du corpus d'apprentissage les plus en rapport avec le document étudié.

Les différentes catégories de connaissances linguistiques que nous avons passées en revue emploient des techniques très différentes, certaines d'entre elles s'intégrant plus ou moins facilement dans un système de RAP. La section suivante décrit quelles sont les adaptations possibles du processus de transcription pour prendre en compte ces nouvelles informations.

2.3 À quel niveau du processus de transcription ?

L'introduction de nouvelles connaissances linguistiques intervient essentiellement au niveau du ML. Parfois, celui-ci est constitué de plusieurs modules ; l'un peut par exemple correspondre à un modèle bigrammes et éliminer avec peu de calculs de nombreuses hypothèses incorrectes, tandis qu'un autre plus complexe apporte des informations supplémentaires sur la syntaxe et la sémantique pour essayer de supprimer les erreurs restantes [Schwartz et Austin, 1990]. La taille de l'espace des hypothèses de transcription tendant à diminuer au fur et à mesure que l'on progresse dans le décodage, chacun de ces modules intervient plus ou moins tôt selon la complexité des calculs demandés. Ceci se manifeste par divers types possibles d'interface entre le MA et le ML. Cette section présente dans un premier temps les modes de couplage MA-ML qui ont été envisagés pour intégrer de nouvelles connaissances linguistiques, notamment les informations syntaxiques qui ont été les plus étudiées. Nous décrivons ensuite les différentes manières de combiner de nouvelles informations avec un modèle N-grammes au sein du ML.

2.3.1 Intégration du modèle acoustique et du modèle de langage

L'intégration du MA et du ML au sein du système de RAP influe principalement sur la complexité des techniques qui peuvent être prises en compte dans le ML. Trois types d'interface entre le MA et le ML peuvent être distinguées : le couplage étroit, le couplage lâche et le couplage modéré [Harper *et al.*, 1994].

Un système à *couplage étroit* considère le MA et le ML comme étant interdépendants et non séparables. Un exemple consiste à représenter le langage d'une application restreinte de dialogue homme-machine par une grammaire à états finis et à incorporer directement cette connaissance dans le MMC du MA. Dans un exemple plus réaliste, un analyseur syntaxique peut être employé pour rechercher les fins possibles de chaque terminal d'une grammaire à partir des vecteurs acoustiques [Ney, 1991]. Ces types de systèmes présentent l'avantage d'utiliser au plus tôt les connaissances linguistiques, essentiellement celles relevant de la syntaxe, mais sont peu utilisés dans la mesure où avoir un seul modèle intégrant le MA et le ML conduit à des solutions algorithmiques de complexité trop importante dans des applications réelles. À cause de l'imbrication du MA et du ML, ils sont en outre difficiles à faire évoluer.

Un système à *couplage lâche* consiste à utiliser le ML pour filtrer ou réordonner les hypothèses fournies par le MA ; de nouvelles connaissances sont alors introduites en post-traitement. Dans le cas où plusieurs informations linguistiques sont utilisées, ce type de couplage permet également d'employer successivement plusieurs modèles linguistiques en faisant en sorte que chacun réordonne la représentation trouvée par le

modèle précédent. Quatre possibilités sont généralement envisagées pour faire l'interface entre les différents modèles : un graphe de mots [Chow et Roukos, 1989 ; Su *et al.*, 1992 ; Harper *et al.*, 1999], une liste des $\mathcal{N}$ meilleures hypothèses [Chow et Schwartz, 1989], un réseau de confusion [Mangu *et al.*, 2000], ou bien encore un graphe d'homophones [Gauvain *et al.*, 2005 ; Béchet *et al.*, 1999]. Si le couplage lâche ne permet pas d'utiliser au plus tôt de nouvelles connaissances pour guider le décodage, il présente l'intérêt de pouvoir employer des méthodes de plus en plus sophistiquées sur un nombre d'hypothèses de plus en plus réduit pour ne pas avoir une trop grande complexité algorithmique. Il offre de plus une grande liberté pour la conception de nouveaux modèles, ce qui permet d'introduire n'importe quel type d'informations linguistiques.

Un système à *couplage modéré* a un comportement intermédiaire entre les deux précédents puisqu'il utilise le ML pour guider le MA, sans y être intégré. Deux approches sont généralement envisagées pour ce type d'interface : descendante et ascendante [Hauenstein et Weber, 1994]. Dans l'*approche descendante*, le ML est invoqué à des points de décision où il prédit des hypothèses. Le MA est ensuite chargé de sélectionner la meilleure hypothèse. Un analyseur syntaxique peut par exemple être utilisé pour prédire des phones qui sont ensuite vérifiés par un MMC de phones [Kita *et al.*, 1989]. Les phones qui constituent un mot sont alors spécifiés par des règles de grammaire. Dans l'*approche ascendante*, les scores acoustiques sont calculés en premier et le ML est appliqué pour vérifier les hypothèses, en réduisant éventuellement le nombre de candidats acoustiques. Cette organisation est très proche de celle des systèmes à couplage lâche mais il est fait appel au ML à chaque point de décision et non pas à la fin de l'analyse du groupe de souffle. Un décodeur tel que nous l'avons décrit à la section 1.1.5 adopte cette approche puisque le ML est utilisé à la fin de chaque hypothèse de mot proposée pour connaître les probabilités conditionnelles sur les mots. Le couplage modéré avec une approche ascendante est actuellement la méthode plus employée pour faire l'interface du MA et du ML dans les systèmes de RAP car elle permet de faire intervenir les informations sur les successions de mots au plus tôt, sans avoir à intégrer directement le ML au MA. Son utilisation impose cependant des contraintes fortes qui sont la rapidité de calcul et la nécessité d'avoir un processus gauche-droit pour le ML, ce que respectent bien les modèles N-grammes. L'usage de l'approche descendante est quant à lui limité dans la mesure où les indices acoustiques semblent pouvoir être plus informatifs que les ML pour éliminer rapidement les hypothèses improbables.

Plusieurs types de couplage peuvent être employés au sein d'un même système de RAP. La plupart des systèmes actuels se trouvent ainsi être multipasses, ce qui permet de recourir au fil des étapes à des MA et des ML de plus en plus informatifs. Un graphe de mots de plus en plus réduit est alors produit à la fin de chaque passe et présenté à la suivante. Le couplage MA-ML au sein d'une passe peut par exemple être modéré, tandis que l'interface entre passes peut être qualifié de lâche.

Après avoir vu dans cette section comment il est possible d'agencer différents modules, nous décrivons dans la suivante comment intégrer de nouvelles informations dans un ML.

2.3.2 Connaissances linguistiques au sein du modèle de langage

Il est possible d'avoir un seul module intégrant plusieurs connaissances linguistiques. Le ML est alors généralement constitué d'un modèle N-grammes, auquel on adjoint des informations supplémentaires selon différents schémas.

Une première solution consiste à construire un nouveau ML plus linguistique que l'on utilise en remplacement d'un modèle N-grammes. Ceci est notamment réalisé dans des systèmes de RAP à langage restreint où une grammaire non contextuelle peut être jugée suffisamment informative pour que l'on n'utilise pas de ML N-grammes [Seneff, 1992]. Parfois, le modèle linguistique inclut aussi déjà des connaissances lexicales sur les probabilités de séquences de mots [Heeman, 1999 ; Wang et Harper, 2002].

Une deuxième possibilité est de continuer à employer un ML N-grammes mais en modifiant ses probabilités à l'aide d'informations plus précises sur un domaine. Ces connaissances peuvent être obtenues à l'aide d'un corpus d'adaptation [Chen *et al.*, 2003] ou bien d'une grammaire non contextuelle probabiliste [Stolcke et Segal, 1994 ; Jurafsky *et al.*, 1995].

Une dernière solution consiste à combiner plusieurs modèles, apportant chacun des connaissances, pour constituer un seul ML. Cette combinaison peut se faire au moyen de l'interpolation linéaire ou du repli, dont les principes ont déjà été exposés (*cf.* section 1.2.1). Dans le cas de l'interpolation linéaire, technique la plus utilisée du fait de sa simplicité, la combinaison de M modèles différents associés aux distributions de probabilité P_k avec $k = 1 \dots M$ s'effectue de la manière suivante :

$$P(w_i|w_1^{i-1}) = \sum_{k=1}^{M} \lambda_k P_k(w_i|w_1^{i-1}) \tag{2.8}$$

où $\sum_{k=1}^{M} = 1$. En ce qui concerne le repli, le calcul pour combiner M modèles, avec M fixé ici à 2 pour simplifier l'équation, s'écrit [Niesler et Woodland, 1996a] :

$$P(w_i|w_1^{i-1}) = \begin{cases} P_1(w_i|w_1^{i-1}) & \text{si } w_i \in \Phi_1(w_1^{i-1}) \\ \alpha(w_1^{i-1}) \times P_2(w_i|w_1^{i-1}) & \text{sinon} \end{cases} \tag{2.9}$$

où $\Phi_1(w_1^{i-1})$ représente l'ensemble des mots dans le contexte w_1^{i-1} pour lequel le 1^{er} modèle est à utiliser en priorité, et α est un coefficient de normalisation. L'intérêt de cette méthode est d'utiliser d'abord les modèles les plus perfectionnés quand on dispose de suffisamment d'informations dans le contexte courant. Une autre possibilité pour combiner des modèles est d'utiliser des *modèles exponentiels*, appelés encore *à entropie maximale* [Rosenfeld, 1996 ; Goodman, 2001]. La plupart des ML sont déterminés en recherchant la distribution de probabilité P qui maximise la probabilité d'observation d'un corpus représentatif du langage à modéliser. Les modèles exponentiels quant à eux se basent sur un critère de maximisation de l'entropie, en recherchant parmi toutes les distributions $\mathcal{P}$ celle qui vérifie :

$$\hat{P} = \arg\max_{P \in \mathcal{P}} H(P) \tag{2.10}$$

où H est la fonction d'entropie. De manière à modéliser les dépendances entre l'historique w_1^{i-1} et le mot à prédire w_i, l'ensemble $\mathcal{P}$ est réduit par des fonctions de contraintes f_k retournant typiquement 0 ou 1. Celles-ci permettent de représenter des modèles N-grammes ou d'autres types de ML ; dans le cas d'une fonction pour le trigrammes $w_a w_b w_c$, on aura ainsi :

$$f_{w_a w_b w_c}(w_1^i) = \begin{cases} 1 & \text{si } w_{i-2} = w_a, w_{i-1} = w_b \text{ et } w_i = w_c \\ 0 & \text{sinon} \end{cases} \tag{2.11}$$

Les modèles satisfaisant le critère du maximum entropie sont dit exponentiels car il peut être montré que leur calculs de probabilités peut être mis sous la forme (*cf.* le tutoriel dans [Ratnaparkhi, 1997]) :

$$P(w_i|w_1^{i-1}) = \frac{\exp(\sum_k \lambda_k f_k(w_1^i))}{Z(w_1^{i-1})} \tag{2.12}$$

où Z est une fonction de normalisation et les λ_k des coefficients obtenus par un algorithme d'apprentissage. Grâce aux fonctions de contraintes, les modèles exponentiels permettent d'intégrer plusieurs sources d'information de manière élégante, comme des connaissances syntaxiques ou thématiques [Wu et Khudanpur, 1999]. Leur temps d'apprentissage est cependant extrêmement long et ils se révèlent assez lents lors de leur utilisation. En outre, hormis avec les modèles à amorces, il semble qu'ils n'aient pas encore permis de réduire la perplexité de manière significative [Goodman, 2001].

2.4 Bilan

Ce chapitre a proposé une synthèse des travaux menés pour introduire des connaissances linguistiques dans le but d'améliorer la transcription. Les deux domaines les plus étudiés sont la syntaxe et l'adaptation des ML. Certaines techniques ont permis une baisse significative du WER ; on peut ainsi citer les ML factorisés pour le cas particulier des langues morphologiquement riches, les analyseurs syntaxiques dans des domaines limités ou sur de la parole lue, l'analyse sémantique latente ou l'utilisation de corpus d'adaptation. Toutefois, malgré des progrès récents, peu de connaissances linguistiques sont encore prises en compte dans les systèmes de RAP. Parmi les raisons qui expliquent cet état de fait, le recours à de nouvelles informations améliore souvent assez peu la qualité de la transcription par rapport aux ML N-grammes, tout en requérant davantage de calculs. En outre, les particularités des transcriptions produites automatiquement, notamment la flexibilité de la langue parlée, la segmentation en groupes de souffle ou encore les erreurs de reconnaissance, viennent compliquer la conception de méthodes automatiques d'acquisition des connaissances linguistiques. De plus, les techniques employées ont souvent les défauts d'augmenter considérablement le temps de décodage du signal acoustique et de ne pas toujours bien s'intégrer dans le cadre statistique des décodeurs de la parole actuels, ce qui fait qu'elles sont utilisées principalement au niveau de la dernière passe du processus de transcription.

Les ML N-grammes, malgré leur indéniable efficacité, restent des modèles assez frustres quant aux connaissances manipulées et des informations linguistiques semblent pertinentes pour rectifier certaines erreurs de transcription. Dans nos travaux, nous avons choisi de nous focaliser sur les informations morpho-syntaxiques et sur l'adaptation thématique. Les raisons de ce choix sont dues aux types d'erreur de transcription que nous avons repérées de la part de notre système de RAP utilisant un ML 4-grammes (*cf.* section 3.2). Nous avons tout d'abord constaté un nombre significatif d'erreurs d'accord, y compris pour des mots consécutifs. Les POS, en introduisant explicitement des informations sur le genre et le nombre, sont adaptées pour corriger des problèmes de ce type. Même si elles ne sont pas à même de détecter aussi finement les relations entre mots distants, elles peuvent être obtenues avec des méthodes plus robustes aux perturbations de la langue orale. Nous nous intéressons au chapitre 4 à la manière dont une transcription peut être étiquetée par les POS, ce qui a été peu étudié jusqu'à présent. Nous présentons ensuite au chapitre 5 un nouveau mode de calcul des probabilités intégrant les informations linguistiques extraites. Le second type de connaissances que nous étudions concerne les thèmes traités par un document. Certaines formules sont spécifiques au sujet abordé ; la suite de mots « *le soleil contesté cet après-midi* » peut ainsi difficilement apparaître dans d'autres contextes que celui de la météo. Le ML que nous employons, construit sur un large corpus très général, gagnerait probablement à être adapté sur des données plus spécialisées. Les chapitres 6, 7 et 8 présentent notre travail exploratoire pour adapter thématiquement le système de transcription. Avant de décrire plus précisément nos travaux, nous exposons dans le chapitre qui suit le cadre expérimental des différents tests menés au cours de notre thèse.

Chapitre 3

Cadre expérimental

Avant de présenter en détail nos contributions, nous exposons le dispositif employé au cours de nos diverses expérimentations. Par souci de clarté et de manière à ne pas disperser les informations que nous pourrions donner à ce sujet, nous avons en effet préféré consacrer un court chapitre à la description des éléments participant aux travaux développés dans la suite de ce mémoire. Nous y présentons tout d'abord le type des documents étudiés, avant de donner les principales caractéristiques du système de RAP que nous utilisons.

3.1 Le corpus Ester

Le corpus employé au cours de nos expérimentations est le corpus Ester. Il a été constitué dans le cadre d'une campagne d'évaluation dédiée à la transcription enrichie et à l'indexation de journaux radiophoniques de langue française [Gravier *et al.*, 2004], et comporte des émissions d'information diffusées par différentes chaînes de radio entre 1998 et 2004. Il correspond essentiellement à de la parole préparée dans la mesure où les principaux intervenants sont des journalistes présentant des titres, des brèves ou des reportages. On trouve toutefois de courts passages de parole spontanée où des personnes sont interrogées pour intervenir sur des points de l'actualité. Certaines émissions peuvent aussi comporter des interviews d'une dizaine de minutes accordées à des personnalités.

Sur l'ensemble du corpus Ester, 100 heures sont transcrites manuellement et accompagnées d'annotations sur la présence de musique, de bruits ou encore d'événements comme les chevauchements de la parole prononcée par plusieurs locuteurs (Fig. 3.1). Chaque émission d'information est généralement subdivisée en entités journalistiques telles que « les titres », « le journal », « la météo », « le point sur le trafic routier »... Toutefois, cette segmentation est très hétérogène puisque certaines émissions sont peu subdivisées, tandis que d'autres contiennent un découpage très précis en associant ainsi une entité à chaque point abordé dans le journal comme « politique internationale - Irak : recherche des objets d'art ». Les émissions sont de surcroît segmentées manuellement en groupes de souffle, en précisant à chaque fois l'identité du locuteur. Ce mode de découpage est normalement effectué en fonction de la durée des

« *Le gouvernement américain va présenter, [i] euh, au conseil de sécurité de l'!ONU, [i] un projet d(e) résolution, [mic] un texte qui permettrait, [i] la levée d(e) l'embargo Noëlle ^^Véli.* »

où « [*i*] » marque une inspiration, « *!ONU* » représente un acronyme, « *d(e)* » indique une absence de prononciation du « *e* », « [*mic*] » témoigne de la présence d'un bruit de micro et « *^^Véli* » précise que le mot est inconnu par le transcripteur.

FIG. 3.1: *Extrait du corpus* ESTER *transcrit selon les conventions de Transcriber*

source	apprentissage		développement
France Inter	19h40	8h	2h
France Info	-	8h	2h
RFI[1]	11h	8h	2h
RTM[2]	-	18h	2h
total	30h40	42h	8h
période	1998-2000	2003	2003

TAB. 3.1: *Répartition du corpus* ESTER *utilisé*

pauses silencieuses. Néanmoins, il semble que certains annotateurs aient également pris en considération des informations d'ordre syntaxique en assimilant par exemple une proposition relative à un groupe de souffle ; cette décision conduit à des unités près de deux fois plus courtes qu'habituellement, peu de groupes de souffle dépassant une taille de dix mots. Conjointement à cette segmentation acoustique, une annotation plus linguistique est proposée en découpant le flux de parole en phrases et en indiquant des signes de ponctuation.

Parmi ces 100 heures d'émissions, seules celles étiquetées dans le corpus comme relevant de l'apprentissage et du développement (Tab. 3.1) ont été utilisées dans nos travaux, ce qui représente au total 1M de mots. La fraction du corpus ESTER destinée à l'apprentissage a permis de construire le MA et le ML du système de RAP employé ; celle dédiée au développement a quant à elle été divisée en deux parts d'une durée de 4 heures chacune. L'une permet de régler des paramètres, notamment la pondération du ML par rapport au MA et la pénalité d'insertion de mot (*cf.* section 1.1.5), tandis que l'autre sert pour les tests du système de RAP (Tab. 3.2).

[1] Radio France internationale.
[2] Radio télévision marocaine.

développement	test
France Inter - 18/04/03 - 7h-8h	France Inter - 18/04/03 - 8h-9h
France Info - 18/04/03 - 12h-13h	France Info - 18/04/03 - 17h-18h
RFI - 08/05/03 - 14h-15h	RFI - 09/05/03 - 14h-15h
RTM - 17/07/03 - 7h-7h15	RTM - 17/07/03 - 13h-13h20
RTM - 17/07/03 - 23h-23h15	RTM - 17/07/03 - 20h-20h20
RTM - 19/07/03 - 7h-7h15	RTM - 19/07/03 - 13h-13h20
RTM - 19/07/03 - 20h-20h15	RTM - 19/07/03 - 23h-23h10
45 565 mots	47 230 mots

TAB. 3.2: *Détail des émissions utilisées comme corpus de développement et de test*

3.2 Le système IRENE

Le système de transcription utilisé au cours de nos expérimentations a été initialement développé par l'IRISA et l'ENST pour la campagne ESTER. Son vocabulaire contient 64 000 mots ; de manière à lui assurer une bonne couverture lexicale, la casse est ignorée et les mots composés[3] se limitent aux plus courants comme « *états-unis* ».

Suivant les principes exposés à la section 1.1, le fonctionnement du système de RAP se décompose en plusieurs étapes (Fig. 3.2). Le signal sonore à analyser est tout d'abord segmenté automatiquement de façon à détecter les zones contenant de la parole. Une subdivision supplémentaire en tours de parole est réalisée au moyen d'un système de reconnaissance du locuteur. Le flux audio est finalement partitionné en groupes de souffle en détectant automatiquement les pauses silencieuses selon la variation de l'énergie. Afin d'éviter des difficultés liées à des erreurs de segmentation, le partitionnement en groupes de souffle que nous avons utilisé est celui réalisé manuellement lors de l'annotation du corpus ESTER. Ce mode de découpage n'introduit toutefois pas un biais important quant à la qualité des résultats produits puisqu'il demeure basé essentiellement sur des informations acoustiques et est beaucoup plus proche d'une segmentation automatique en groupes de souffle que d'une segmentation linguistique s'appuyant sur les signes de ponctuations annotés.

Chaque groupe de souffle est analysé séparément par un système à trois passes. Une première passe, utilisant un MA assez simple non contextuel[4] et un ML trigrammes de mots, vise à produire des graphes de mots de taille conséquente. Ces graphes sont ensuite réévalués dans une seconde passe par un MA plus complexe dépendant du contexte[5] et un ML quadigrammes de mots. Une dernière passe, basée sur la transcription obtenue à l'issue de l'étape précédente et la partition en locuteurs, utilise des MA adaptés selon chaque locuteur suivant la méthode MLLR (*Maximum Likelihood Linear Regression*). Cette technique vise à transformer linéairement les gaussiennes du MA en les adaptant à

[3] Mots contenant des « - » ou « ' ».
[4] Modélisant des phones de manière isolée.
[5] Modélisant des triphones.

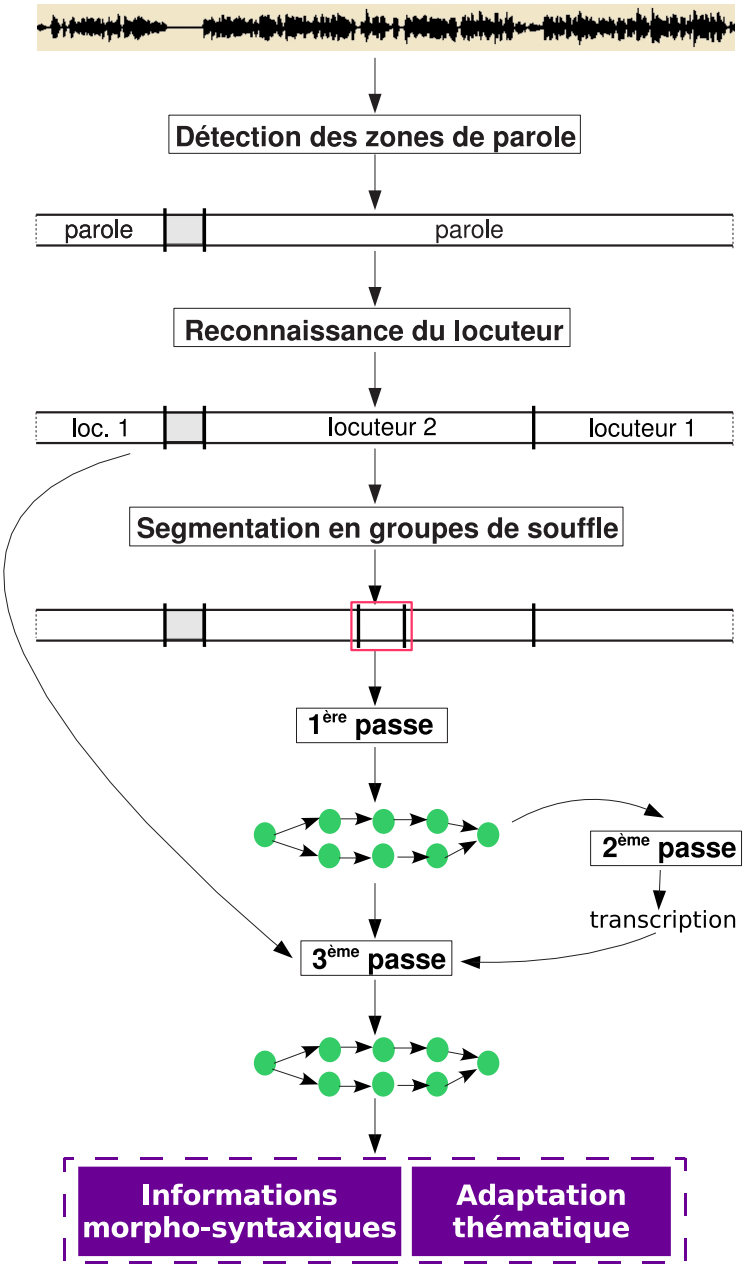

FIG. 3.2: *Constituants du système* IRENE

de petit volumes de données propres à chaque locuteur [Leggetter et Woodland, 1995]. Le ML utilisé dans la dernière passe est le même que celui utilisé dans la précédente. Ce ML quadrigrammes, de même que le ML trigrammes de la première passe, sont estimés en interpolant les probabilités calculées sur 350 M de mots du journal *Le Monde* avec celles déterminées sur les 1 M de mots de l'extrait du corpus ESTER destiné à l'apprentissage (Tab. 3.1).

Ce sont les graphes de mots produits par la troisième passe que nous utilisons au cours de nos expérimentations. Ces graphes permettent notamment de produire les listes des $\mathcal{N}$ meilleures hypothèses en extrayant les chemins qui obtiennent les meilleurs scores (*cf.* équation (1.11) page 16) et en ne retenant que des hypothèses qui diffèrent d'au moins un mot. Nous avons recouru à deux valeurs différentes de $\mathcal{N}$ au cours de nos expériences. Les listes correspondant à $\mathcal{N} = 100$ ont été obtenues en limitant les graphes de mots produits à la première passe aux 1000 meilleurs chemins ; celles correspondant à $\mathcal{N} = 1000$ ont été construites sans restriction sur la taille de ces graphes.

Le système de RAP employé étant décrit, nous présentons dans la partie suivante de ce mémoire nos travaux sur l'utilisation des POS en RAP, en débutant par un chapitre consacré plus particulièrement à l'étiquetage de transcriptions.

Deuxième partie

Informations morpho-syntaxiques et reconnaissance de la parole

Chapitre 4

Étiquetage morpho-syntaxique de l'oral

Les étiquettes POS présentent deux intérêts pour la RAP. Elles constituent d'une part une information *a priori* intéressante pour corriger des erreurs de transcription puisque par rapport aux ML opérant uniquement sur des mots, elles apportent des connaissances explicites sur les catégories grammaticales présentes dans un groupe de souffle, ce qui permet d'introduire des contraintes simples sur la syntaxe. De plus, si elles sont accompagnées d'informations sur le genre et le nombre, elles peuvent jouer un rôle important dans la prise en compte des accords. D'autre part, trouver les étiquettes POS associées à chaque mot d'une transcription est une étape nécessaire à de nombreux traitements. La construction d'un index à partir de documents sonores — pour ne citer qu'un exemple — peut ainsi être réalisée en ne conservant que les noms communs.

Les premiers pas de l'utilisation des POS en reconnaissance de la parole (*cf.* section 2.2.3.1) remontent aux années 80 et au début des années 90 [Jelinek, 1990 ; Maltese et Mancini, 1992] ; cette prise en compte est alors intervenue au moyen des modèles N-classes. Cette méthode, consistant à réaliser une interpolation linéaire avec des ML N-grammes de mots, n'apporte généralement pas de gain au niveau de la perplexité et du WER [Weintraub *et al.*, 1996], ce qui fait que l'on privilégie plutôt des classes basées sur des critères statistiques [Niesler *et al.*, 1998]. Heeman [1999] propose un nouveau mode d'intégration des POS en utilisant un ML trigrammes de paires mots/POS plutôt que de mots et en redéfinissant le problème de la RAP comme une recherche de la meilleure séquence conjointe de mots et de POS. Cette technique conduit à une baisse significative du WER mais augmente fortement le nombre d'entrées du ML, ce qui requiert de très nombreuses données d'apprentissage et repose lourdement sur des techniques de lissage. Les travaux s'intéressant spécifiquement aux caractéristiques sur le genre et le nombre restent en revanche peu nombreux. Lavecchia *et al.* [2006] présentent toutefois un modèle à base de cache utilisant cette connaissance en observant sur le français une amélioration de la qualité de la transcription.

Dans ce mémoire, nous envisageons les POS en RAP avec une approche différente puisque nous cherchons dans un premier temps à étiqueter les hypothèses de mots produites par leur classe grammaticale, avant d'employer cette connaissance pour améliorer la RAP. Nous introduisons en outre explicitement des informations sur le genre et le nombre, tout en prenant en compte l'ordre des séquences de POS trouvées, à la différence des modèles recourant à des mémoires caches. Une autre originalité réside dans notre mode intégration des POS qui se fait au niveau du calcul des probabilités associées à un groupe de souffle et non à un mot en particulier. Enfin, contrairement à la plupart des études que nous avons pu trouver à ce sujet, nous testons notre méthode avec un ML quadigrammes appris sur un corpus de grande taille, et donc *a priori* déjà assez performant pour modéliser le langage.

La première partie de notre travail a consisté à déterminer la proportion d'erreurs de transcription corrigeables par la connaissance des POS. Notre objectif était alors de savoir si cette information présentait un intérêt pour améliorer la qualité des résultats produits par un système de RAP. La réponse à cette question étant positive, nous avons dans une deuxième partie étudié le comportement des étiqueteurs morpho-syntaxiques sur une transcription. Ces outils ont fait l'objet de nombreuses recherches dans le cadre d'annotation automatique de corpus et il en existe actuellement plusieurs versions opérationnelles sur des textes de la langue écrite. Toutefois, ils ont été très peu étudiés sur des documents oraux, et notamment sur des transcriptions produites par des systèmes de RAP, alors que les spécificités de la langue parlée et les erreurs de transcription sont sujettes à complexifier l'étiquetage automatique. Nous avons donc mené différentes évaluations destinées à montrer l'aptitude des étiqueteurs à traiter des documents oraux, en prévision d'une utilisation des POS en post-traitement d'un système de transcription.

Avant de voir spécifiquement dans le chapitre 5 notre emploi des POS dans un système de RAP, ce chapitre décrit comment obtenir un étiquetage automatique d'une transcription. Nous établissons tout d'abord une typologie des erreurs de transcription, en mettant en exergue celles qui sont *a priori* détectables par des informations morpho-syntaxiques. Ayant été amené à développer notre propre étiqueteur pour avoir plus de flexibilité sur la prise en compte des POS obtenues dans le système de RAP, nous donnons les principes de l'étiquetage catégoriel, puis faisons une description de notre outil. Nous concluons le chapitre par une évaluation des étiqueteurs que nous avons testés sur l'oral.

4.1 Typologie des erreurs de transcription

Afin d'évaluer l'apport potentiel des POS pour la transcription, nous avons étudié en détail un court extrait des transcriptions automatiques produites par le système IRENE, en cherchant à connaître la part des erreurs corrigeables par cette seule connaissance. Cet extrait, constitué de 6 500 mots issus du corpus de développement, soit 563 groupes de souffle, correspond à un peu plus de 30 minutes d'une émission d'information diffusée sur France Inter. Le WER sur cet extrait est de 17,8%. Parmi les erreurs de

reconnaissance que nous y avons constatées manuellement, trois groupes principaux se détachent.

Certaines erreurs correspondent à un « dérapage » du système, généralement dû soit à une mauvaise acoustique, soit à une mauvaise reconnaissance d'entités nommées, notamment des noms de personnes (Fig. 4.2, deux premiers exemples). Deux extraits ayant un bruit de fond gênant pour la compréhension et représentant 19 groupes de souffle, soit moins de deux minutes au total, sont ainsi tout de même responsables de 15,3 % des mots mal reconnus. Les erreurs sur la reconnaissance des entités nommées et de certains termes techniques, tel que « *psycho-oncologie* », représentent quant à elles 11,3% du nombre d'erreurs de transcription. Il est à noter que l'extrait étudié a été diffusé la même semaine que certaines émissions dont fait partie le corpus d'apprentissage d'ESTER et à partir duquel a été construit le dictionnaire de prononciation. Il est certain que la part de ce type d'erreurs aurait augmenté si le document à transcrire était produit dans un temps plus lointain, avec des sujets d'actualité qui se seraient depuis renouvelés. Une autre configuration où le système de RAP produit consécutivement de nombreuses erreurs concerne les groupes de souffle très courts ; la récupération d'une mauvaise reconnaissance acoustique par le ML est alors rendue difficile du fait de la taille limitée du contexte. Afin d'illustrer l'influence de la longueur du groupe de souffle sur le nombre d'erreurs, nous donnons sur la figure 4.1 la mesure du WER calculée en fonction de cette variable, sur un corpus étendu représentant 4 heures d'émission d'information. Ce graphique montre une augmentation significative du nombre d'erreurs lorsque les groupes de souffle sont constitués d'au plus 5 mots. Sur l'échantillon analysé manuellement de 6 500 mots, les groupes de souffle de cette taille représentent toutefois une part assez faible (4,4 % des mots et 5,2% des erreurs) des mots mal transcrits. L'ensemble de ces erreurs se produisant « en rafales » semblent en tout cas hors d'atteinte de la correction susceptible d'être apportée par les POS.

Le deuxième groupe que nous avons identifié correspond à des hypothèses de groupes de souffle transcrites avec des agrammaticalités (Fig. 4.2, troisième exemple). Ces dernières sont notamment engendrées par des mots grammaticaux courts, notamment les auxiliaires « *a* », « *ont* » et « *est* », les prépositions « *dans* », « *en* », « *de* » et « *à* », les conjonctions « *et* », « *ou* » et « *que* », ainsi que les déterminants « *les* », « *des* », « *ces* » et « *ses* ». Ces mots absents ou présents de manière inopinée dans les hypothèses de transcriptions sont particulièrement fréquentes puisque plus d'un groupe de souffle sur cinq contient une erreur sur ce type de mots. Outre ces fautes, on retrouve également des mots mal reconnus dus à des confusions sur le temps et le mode des verbes, le présent et l'indicatif étant souvent privilégiés ; ce type d'erreurs représente 2,4 % de l'ensemble de celles observées sur le corpus étudié. Une autre explication de la présence d'hypothèses agrammaticales est directement lié au mode de segmentation qui n'est pas assez linguistique. Ainsi avons-nous remarqué que les groupes de souffle débutant par des verbes étaient souvent sujets à des erreurs de transcription (Fig. 4.2, quatrième exemple). Nous avons enfin pu identifier une dernière source de production d'hypothèses non grammaticales, à savoir la présence de répétitions et de reprises dans le flux de parole. Ces erreurs constituent une part relative de 2,6 %, chiffre qui serait bien entendu plus important si l'extrait étudié contenait davantage de parole spontanée. Parmi toutes

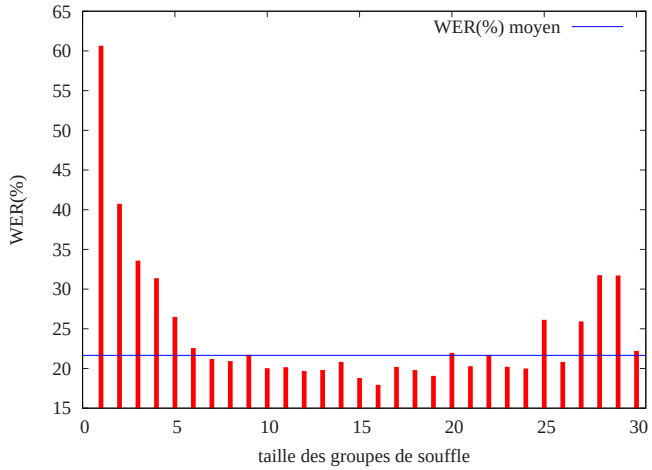

FIG. 4.1: *Évolution du WER (%) en fonction de la taille des groupes de souffle exprimée en nombre de mots de la référence. Les barres verticales ont été produites en divisant le corpus en échantillons contenant des groupes de souffle de même taille. Les valeurs données sur ce graphique sont établies sur un ensemble d'émissions de 4 heures, dont est extrait le corpus étudié de 6 500 mots et sur lequel est observé un WER égal à 21,7 %. Les groupes de souffle d'une taille supérieure à 30 mots ont été écartés car ils étaient très peu nombreux (0,2 % de l'ensemble des groupes de souffle)*

les agrammaticalités observées, certaines semblent corrigeables puisque l'étiquetage des groupes de souffle peut conduire à des séquences de POS aberrantes, comme l'apparition de trois prépositions consécutives. Ce critère est néanmoins à prendre avec précaution, à cause des répétitions présentes dans la langue parlée. En outre, les POS ne peuvent prétendre à elles seules corriger des erreurs sur la syntaxe.

Le troisième groupe est formé d'erreurs très vraisemblablement corrigeables grâce aux POS, à savoir les fautes d'accord en genre et en nombre et les confusions entre infinitif et participe passé. Ces erreurs sont particulièrement nombreuses puisqu'elles représentent 11,7 % des erreurs. Parmi elles, certaines nécessitent une résolution d'anaphores (2,9 % du nombre total des mots mal reconnus, Fig. 4.2, cinquième exemple), le référant étant dans un autre groupe de souffle ; d'autres (1,3 %, Fig. 4.2, sixième exemple) correspondent à des accords avec une entité qui est absente du groupe de souffle contenant le verbe. Notons que ces erreurs sont inaccessibles en analysant isolément chaque groupe de souffle. Parmi les fautes d'accord, on retrouve de plus des choix tout à fait valables effectués par le système de RAP mais pourtant considérés comme erronés (1,0 %, Fig. 4.2, septième exemple). Ceci est expliqué par le fait que dans certaines configurations, plusieurs accords sont possibles alors qu'il n'y en a qu'une seule qui est proposée dans la transcription de référence[1]. Nous avons finalement dénombré 76 erreurs (6,5 % des mots mal reconnus) rectifiables sans avoir à examiner de dépendances entre des groupes de souffle consécutifs (Fig. 4.2, huitième exemple), dont 5 (0,4 %) se

[1] Il faudrait indiquer plusieurs transcriptions de référence pour évaluer très précisément le WER. Les cas de figures où ceci s'avère nécessaire restent cependant rare.

1 - Entités nommées mal transcrites
REF: il y avait ********* ÉGALEMENT BARZAN AL TIKRITI en fait les principaux ET également TAREK AZIZ PAR exemple
HYP: il y avait TELLEMENT BAS LA PARTIE CRITIQUE en fait les principaux ** également ***** ARRÊTÉS HISTOIRE exemple
2 - Groupe de souffle court mal transcrit
REF: DE L' AMENDE
HYP: ** TELLEMENT MÊME
3 - Hypothèse agrammaticale
REF: bush ** SAIT donc QU' il faudra coopérer
HYP: bush S' EST donc ** il faudra coopérer
4 - Erreur sur un groupe de souffle débutant par un verbe
REF: ** FAIT aussi pâle figure ou presque
HYP: C' EST aussi pâle figure ou presque
5 - Accord nécessitant une résolution d'anaphore
REF: ILS POURRAIENT être ANNONCÉS début juin
HYP: IL POURRAIT être ANNONCÉ début juin
6 - Accord avec un sujet absent du groupe de souffle
REF: FERMENT nos usines
HYP: FERME nos usines
7 - Accord acceptable détectée comme une erreur
REF: l' ensemble des personnes qui APPARTENAIENT au au parti baath
HYP: l' ensemble des personnes qui APPARTENAIT au au parti baath
8 - Accord interne au groupe de souffle
REF: c' est un monstre injuste envers sa soeur si DÉVOUÉE
HYP: c' est un monstre injuste envers sa soeur si DÉVOUÉ
9 - Confusion entre des quasi-homophones
REF: c' est la pathologie de L' INCERTITUDE
HYP: c' est la pathologie de LA CERTITUDE

FIG. 4.2: *Exemples d'erreurs dans les groupes de souffle*

produisent sur des mots pour lesquels la flexion correcte est absente du dictionnaire du système de RAP. Les erreurs corrigeables sans augmenter la taille du vocabulaire et en se restreignant à une analyse indépendante de chaque groupe de souffle représente donc 6,1 % des mots mal reconnus et les corriger représenterait une baisse absolue de 1,1 % du WER.

Au travers de l'exposé des principales erreurs de décodage, nous avons distingué trois principales catégories. La classification est arbitraire, les erreurs d'accord pouvant être par exemple considérées comme des agrammaticalités, de même que les hypo-

thèses contenant de nombreux mots mal reconnus. Elle n'inclut pas toutes les erreurs, la correction de certaines nécessitant par exemple des connaissances d'ordre sémantique ou contextuelle (confusion entre « *l' incertitude* » et « *la certitude* », Fig. 4.2, dernier exemple). D'une manière générale, l'essentiel des erreurs de transcription se produisent sur des mots monosyllabiques, notamment les mots grammaticaux, ou sur l'emploi d'un quasi-homophone à la place d'un autre, comme « *c'est un ordre* », au lieu de la séquence réellement prononcée « *c'est un ogre* ». Malgré les difficultés que pose la réalisation d'une telle classification des erreurs, il apparaît toutefois qu'un nombre significatif de mots mal reconnus est lié à une mauvaise prise en compte des phénomènes d'accord, ce qui positionne les POS comme une source d'information susceptible d'être intéressante pour améliorer la qualité de la transcription. La particularité du corpus étudié qui explique le nombre assez élevé de fautes d'accord est qu'il est de langue française ; il est probable que pour d'autres langues possédant moins de flexions homophones comme l'anglais les observations auraient différé. Une étude a montré à ce sujet que 23 % des mots d'un extrait du corpus du *Wall Street Journal* possédaient au moins un homophone, alors que ce chiffre s'élevait à 75 % sur un extrait du corpus *Le Monde* [Gauvain *et al.*, 1994].

L'analyse des erreurs de transcription ayant établi l'intérêt des POS, l'étape suivante de notre travail a consisté à examiner si nous pouvions obtenir un étiquetage fiable de transcriptions automatiques.

4.2 Principes de l'étiquetage morpho-syntaxique

L'étiquetage d'un texte consiste à assigner à chaque mot w_i d'une phrase $W = w_1 \dots w_n$ une classe t_i parmi un ensemble fini d'étiquettes préalablement déterminées. Dans un premier temps, le recours à un dictionnaire ou une analyse morphologique détermine les catégories possibles de chaque mot. Des règles sur les suffixes sont en particulier très utiles pour attribuer des étiquettes aux mots inconnus. Les ambiguïtés sont ensuite levées en sélectionnant une seule séquence $T = t_1 \dots t_n$ que l'on espère correcte du point de vue grammatical, bien qu'il soit possible dans certains cas d'avoir plusieurs désambiguïsations correctes.

Quelle que soit la méthode utilisée, les étiqueteurs morpho-syntaxiques sont reconnus à l'heure actuelle comme ayant des performances très satisfaisantes [Véronis, 2000]. Les résultats publiés concernant les documents écrits sont couramment de l'ordre de 95 % de classes catégorielles bien attribuées et l'on trouve même des nombres supérieurs. Les sections suivantes font un rapide bilan des techniques employées pour étiqueter, avant de présenter les premières études qui ont été faites concernant l'oral.

4.2.1 Techniques utilisées

Les techniques couramment utilisées aujourd'hui pour étiqueter un texte reposent toutes sur l'examen statistique des séquences de mots et de POS, en faisant abstraction de toute analyse syntaxique ou du discours. Une des grandes avancées dans ce domaine a d'ailleurs consisté à établir qu'elles permettaient d'obtenir les POS de manière fiable,

alors que l'étiquetage était auparavant souvent considéré comme faisant partie du problème plus général de la compréhension de la langue naturelle [Abney, 1996]. Toutefois, même si les 95 % mentionnés pour les performances semblent élevés, il convient de le relativiser. En effet, de nombreux mots ne sont pas ambigus et lorsque plusieurs étiquettes sont possibles, le choix de l'étiquette la plus fréquente permet déjà de désambiguïser correctement une part importante des mots d'un texte, notamment lorsque les textes ayant permis de comptabiliser les occurrences sont issus de domaines proches de ceux traités par les documents à étiqueter. Ainsi, une étude sur un corpus du journal *Le Monde* a montré que sur ce simple principe 92,1 % des étiquettes sont correctes au niveau des catégories de base (nom, verbe, *etc.*) et que 82,5 % le sont quand on utilise des étiquettes plus fines accompagnées d'informations morphologiques [Vergne, 1999]. L'objectif des méthodes plus sophistiquées va donc consister à améliorer cette technique rudimentaire.

Un premier type d'étiqueteurs repose sur l'utilisation de *règles*. La désambiguïsation s'effectue alors en assignant un ensemble d'étiquettes possibles sur la base d'un lexique ou d'une analyse morphologique puis à ne retenir qu'une seule étiquette pour chaque mot au moyen de règles contextuelles telles que : « le mot courant n'est pas un verbe si le mot précédent est un déterminant ». Même si leurs performances sont aussi bonnes voire meilleures que celles obtenues avec les autres types d'étiqueteurs, une critique qui leur est souvent faite est que l'écriture manuelle des règles de désambiguïsation demande un effort important. Il est toutefois possible d'avoir recours à des méthodes automatiques. L'étiqueteur de Brill permet ainsi d'acquérir mécaniquement des règles à partir d'un corpus d'apprentissage déjà étiqueté manuellement [Brill, 1992; 1994].

Un deuxième ensemble d'étiqueteurs correspond à des *modèles stochastiques*. Une technique très utilisée repose sur les MMC [Merialdo, 1994], la modélisation s'effectuant alors en considérant que les variables cachées sont des étiquettes et les observations les mots à étiqueter. Elle suppose en outre, afin de limiter le nombre de paramètres à estimer, que la probabilité d'avoir une étiquette dépende uniquement du mot à annoter et des $N-1$ étiquettes précédentes. La phase d'étiquetage revient ainsi à rechercher la succession d'étiquettes la plus probable $\hat{T}$ pour la suite de mots W en supposant :

$$P(W,T) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-N+1}\dots t_{i-1}) \ . \tag{4.1}$$

Remarquons au passage que cette égalité se rapproche de celle utilisée par les modèles N-classes (*cf.* équation (1.20) page 22). En pratique, N est souvent limité à 3 [Merialdo, 1994 ; Brants, 2000], les quadrigrammes augmentant considérablement la taille du MMC tout en n'améliorant pas significativement la qualité de l'étiquetage. L'étiqueteur MULTEXT [Armstrong *et al.*, 1995] basé sur les MMC se limite même à $N = 2$. Il présente en outre la particularité de rassembler au sein d'une même classe les mots possédant d'après le lexique le même ensemble d'étiquettes ambiguës ; une observation se retrouve alors être une classe plutôt qu'un mot particulier, ce qui permet de réduire la taille du MMC.

D'autres méthodes statistiques sont également utilisées, donnant des résultats tout aussi satisfaisants. On peut ainsi citer l'étiqueteur TREETAGGER, qui détermine la

séquence d'étiquette la plus probable de manière similaire au modèle précédent à la différence près que les probabilités $P(t_i|t_{i-N+1}^{i-1})$ sont déterminées à l'aide d'un arbre de décision binaire et non plus d'un MMC [Schmid, 1994; 1995]. Des étiqueteurs ont également été construits en utilisant des modèles à maximum d'entropie [Ratnaparkhi, 1996] ou des machines à vecteurs de support (SVM pour *Support Vector Machine*) [Giménez et Màrquez, 2004]. Les méthodes à base de MMC, bien que très simples, ont montré des résultats comparables à celles obtenues avec ces techniques plus complexes [Brants, 2000]. Nous utilisons ainsi lors de la conception de notre étiqueteur un MMC d'ordre 3.

Tous ces outils sont généralement appliqués sur des textes de langue écrite. Depuis peu, certains travaux sont réalisés pour étudier leur comportement sur des documents oraux ; la section suivante se propose d'en faire un bilan.

4.2.2 Étiquetage de documents oraux

Si l'étiquetage des documents écrits par des catégories grammaticales a débuté dès les années 60 [Véronis, 2000], ce n'est que récemment que des études ont été menées sur les documents oraux et elles ne sont pas encore très nombreuses. La constitution d'étiqueteurs pour l'oral se heurte encore à l'absence de corpus oraux étiquetés préexistants. Les premières expérimentations ont donc été menées sur la langue parlée à partir d'étiqueteurs conçus initialement pour les documents écrits, que ce soit pour l'anglais [Garside, 1995], le français [Valli et Véronis, 1999 ; Campione *et al.*, 2005], le suédois [Nivre et Grönqvist, 2001], le néerlandais [Van Eynde *et al.*, 2000], l'espagnol [Moreno et Guirao, 2003], le portugais [Mendes *et al.*, 2003], l'italien [Panunzi *et al.*, 2004] ou encore le japonais [Uchimoto *et al.*, 2002]. Le processus d'étiquetage subit souvent quelques adaptations à l'oral. Ces modifications se présentent sous la forme d'un prétraitement du corpus à étiqueter, ou d'un changement du comportement de l'étiqueteur pour des mots spécifiques. Parmi les opérations de prétraitement, les pauses silencieuses, qui permettent une segmentation plus fine que celle des tours de parole, peuvent ainsi être remplacées par des points de suspension, *i.e.*, par la ponctuation la plus neutre possible par rapport au fonctionnement de l'étiqueteur. Les amorces de mots sont quant à elles généralement supprimées automatiquement car il est parfois difficile de deviner le mot prononcé. Il peut être également utile d'éliminer des événements non lexicaux, tels que « *hein* » et « *euh* ». De plus, certains mots apparaissent beaucoup plus fréquemment à l'oral qu'à l'écrit, notamment les contractions de mots et les interjections ; les mots caractéristiques de l'oral sont ainsi insérés dans le lexique de l'étiqueteur. De même, des règles peuvent être manuellement ajoutées et des probabilités d'assignation des étiquettes modifiées pour les mots qui ont un comportement différent à l'écrit et à l'oral.

Les résultats publiés concernant l'étiquetage morpho-syntaxique de documents oraux se rapprochent des performances déjà observées sur l'écrit puisque le pourcentage d'étiquettes correctes atteint 97,9 % pour le français [Valli et Véronis, 1999], entre 95 % et 97 % pour le suédois [Nivre et Grönqvist, 2001], 94,3 % pour le néerlandais [Van Eynde *et al.*, 2000] ou encore 98,3 % pour l'espagnol [Moreno et Guirao, 2003].

Ces résultats sont difficilement comparables puisque les jeux d'étiquettes sont de granularités diverses et concernent des langues différentes. Ils montrent tous cependant la capacité des étiqueteurs conçus pour l'écrit à s'adapter aux spécificités de la langue parlée, notamment aux disfluences. La raison qui est souvent invoquée pour expliquer cet état de fait est que les outils utilisés prennent en compte un contexte limité à des trigrammes, ce qui autorise une certaine flexibilité lors de la désambiguïsation.

Si des études apparaissent depuis quelque temps sur l'étiquetage de documents relevant de la langue orale, le comportement des étiqueteurs morpho-syntaxiques sur des transcriptions produites par un système de RAP — sans indices sur la ponctuation et avec des mots erronés — reste encore mal connu. La section suivante présente un étiqueteur statistique que nous avons développé pour l'oral et que nous évaluons par la suite sur des transcriptions automatiques.

4.3 Conception d'un étiqueteur morpho-syntaxique pour l'oral

Il existe de nombreux étiqueteurs morpho-syntaxiques disponibles ; nous avons toutefois décidé de développer notre propre étiqueteur. Les raisons sont nombreuses et essentiellement pratiques. Cela nous permet tout d'abord de pouvoir modifier plus facilement notre outil pour le paramétrer pour l'oral, les étiqueteurs n'étant pas toujours distribués en compagnie de leur code. De plus, nous voulons un outil assez rapide pour étiqueter un grand nombre d'hypothèses de transcription en un temps raisonnable. L'étiqueteur MULTEXT, lent lors de l'étiquetage, ne satisfaisait pas à ce critère. En outre, nous souhaitons un étiqueteur suffisamment flexible pour choisir nos propres jeux d'étiquettes et le vocabulaire des mots reconnus par l'étiqueteur. Cette dernière possibilité s'est avéré importante dans la mesure où les noms propres, transcrits sans majuscules par notre système de RAP, sont systématiquement associés à une mauvaise classe grammaticale pour la plupart des outils que nous avons pu tester.

Nous avons choisi de construire un étiqueteur stochastique basé sur les MMC, une première comparaison de la qualité de l'étiquetage produit par différents outils sur des transcriptions automatiques montrant que l'étiqueteur de Brill conduisait à des résultats un peu moins bons que ceux obtenus avec MULTEXT, TNT [Brants, 2000] ou TREETAGGER. L'utilisation d'étiqueteurs statistiques présente l'inconvénient de devoir construire un corpus étiqueté, mais ce corpus nous est également utile pour apprendre les paramètres du modèle intégrant les POS dans le calcul des probabilités du système de RAP (*cf.* section 5.2). La suite de cette section expose le modèle statistique qui est à la base de notre étiqueteur, puis présente brièvement le corpus d'apprentissage et le lexique que nous avons utilisés pour notre étiqueteur. Elle décrit ensuite la phase d'apprentissage de l'étiqueteur, puis celle d'étiquetage.

4.3.1 Le modèle utilisé et ses paramètres

Notre étiqueteur évalue la probabilité que W soit associé à la suite d'étiquettes T à partir des paramètres $P(w_i|t_i)$ et $P(t_i|t_{i-N+1}^{i-1})$ (*cf.* équation (4.1)). Il existe deux approches pour estimer ces deux types de paramètres [Merialdo, 1994]. Une première consiste à recourir à l'algorithme ordinairement utilisé pour l'apprentissage des probabilités dans un MMC, *i.e.*, l'algorithme de Baum-Welch. Sa particularité consiste à estimer simultanément les deux types de paramètres en cherchant le modèle $\hat{\mathcal{M}}$ qui satisfait au critère du maximum de vraisemblance :

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \sum_{T} P_{\mathcal{M}}(W,T) \tag{4.2}$$

où W représente les mots du corpus et T une succession possible d'étiquettes pour W. Le principal avantage de cette méthode est de pouvoir paramétrer le modèle à partir d'un corpus non étiqueté en estimant de manière itérative les probabilités, jusqu'à ce que la probabilité $\sum_T P_{\mathcal{M}}(W,T)$ n'augmente plus. Toutefois, pour obtenir de meilleurs résultats, l'algorithme gagne à être initialisé à partir des nombres d'occurrences des séquences t_{i-N+1}^{i} dans un corpus étiqueté, ce qui réduit son intérêt.

Une seconde approche évalue séparément les deux types de probabilité $P(w_i|t_i)$ et $P(t_i|t_{i-N+1}^{i-1})$ au moyen des fréquences relatives. Les valeurs sont alors déterminées à partir des nombres d'occurrences des étiquettes et des mots dans un corpus d'apprentissage :

$$P(w_i|t_i) = \frac{C(w_i,t_i)}{C(t_i)} \tag{4.3}$$

$$P(t_i|t_{i-N+1}^{i-1}) = \frac{C(t_{i-N+1}^{i})}{C(t_{i-N+1}^{i-1})} \tag{4.4}$$

où $C(w_i,t_i)$ représente le nombre de fois où le mot w_i est associé à l'étiquette t_i et $C(t_{i-N+1}^{i})$ le nombre d'occurrences de la séquence d'étiquettes t_{i-N+1}^{i} dans un corpus d'apprentissage étiqueté. Même si les deux types de paramètres ne sont pas réestimés ensemble pour améliorer la perplexité (*cf.* section 1.3.1) du corpus d'apprentissage comme dans le cas de l'algorithme de Baum-Welch, cette seconde approche a des performances d'étiquetage aussi bonnes, voire meilleures que la première [Merialdo, 1994]. Nous avons donc opté pour un apprentissage basé sur les fréquences relatives.

4.3.2 Corpus d'apprentissage et vocabulaire

Le corpus utilisé pour l'apprentissage des paramètres de l'étiqueteur est un extrait du corpus Ester. Ce corpus contenant des émissions d'information radiophoniques, il relève à la fois de la parole préparée et spontanée et est donc plus adéquat que des corpus de langue écrite pour étiqueter des documents oraux. Une expérience menée à partir de deux étiqueteurs[2] a montré l'intérêt d'estimer les paramètres sur un corpus de langue parlée pour étiqueter de l'oral puisqu'en faisant l'apprentissage sur un extrait du

[2] L'étiqueteur de Brill et un étiqueteur basé sur le critère du maximum d'entropie.

corpus oral SWITCHBOARD, le pourcentage d'étiquettes correctes avoisinait 85 %, alors qu'il n'était que de 75 % à partir du *Wall Street Journal* [Hamaker, 1999].

Nous avons constitué un corpus d'apprentissage d'environ 200 000 mots représentant une durée de 16 heures. Il est extrait de la partie du corpus ESTER que nous avons utilisée pour l'apprentissage du système de RAP (Tab. 3.1 page 54). Les transcriptions manuelles, contenant à l'origine des majuscules et des ponctuations, ont été étiquetées par des POS au moyen du logiciel CORDIAL[3]. Le résultat a été converti automatiquement vers le jeu d'étiquettes que nous avons adopté, puis vérifié manuellement pour corriger les erreurs d'étiquetage. Conservées dans un premier temps pour ne pas perturber CORDIAL, les majuscules et les marques de ponctuation dans le texte ont ensuite été ôtées, avant de finalement segmenter le résultat en groupes de souffle pour que le format de notre corpus soit cohérent avec celui du texte produit par notre système de transcription.

Nous recourons à un lexique étiqueté afin de connaître les POS possibles pour chaque mot. Ce dictionnaire a été construit à partir de celui élaboré pour le français dans le cadre du projet MULTEXT [Ide et Véronis, 1994], en le restreignant aux mots appartenant au dictionnaire de prononciation que nous utilisons[4]. Il ne contient ni ponctuation, ni mots composés. Nous avons toutefois ajouté une dizaine de locutions telles que « *parce que* » ou « *a priori* ». Le choix des étiquettes morpho-syntaxiques a été fait de manière à discriminer le genre et le nombre des adjectifs, des noms, des déterminants et des pronoms, et le temps et le mode des verbes, ce qui conduit à un jeu de 93 étiquettes différentes (*cf.* annexe A). Cet ensemble d'étiquettes est très proche de celui proposé dans les grammaires scolaires et est directement inspiré de celui de CORDIAL.

4.3.3 Apprentissage du modèle

L'apprentissage de l'étiqueteur consiste à estimer deux types de paramètres : les probabilités lexicales $P(w_i|t_i)$ et les probabilités contextuelles $P(t_i|t_{i-N+1}^{i-1})$.

4.3.3.1 Probabilités lexicales

Un vocabulaire ayant été fixé pour l'étiqueteur, seules les probabilités lexicales $P(w_i|t_i)$ concernant les mots w_i de ce vocabulaire sont calculées à partir du corpus d'apprentissage. En outre, l'application que nous visons étant l'étiquetage de transcriptions automatiques produites par notre propre système de RAP, les textes à étiqueter ne peuvent pas contenir de mots hors vocabulaire. Il n'est donc pas nécessaire d'effectuer un traitement particulier pour les mots inconnus en réservant par exemple une masse de probabilité $P(w_{\text{unk}}|t_i)$ aux mots absents du vocabulaire du système de RAP mais susceptibles de se présenter dans les textes à étiqueter. Certains mots du dictionnaire sont en revanche absents du corpus d'apprentissage, ce qui conduit à adopter la stratégie suivante. Dans le cas où un mot w_i est associé une étiquette t_i totalement absente

[3] Version 8.1 distribuée par la société Synapse Développement.

[4] Nous tenons à remercier François Yvon pour nous avoir fourni le dictionnaire étiqueté.

du corpus d'apprentissage[5], la probabilité $P(w_i|t_i)$ est estimée selon une distribution répartie uniformément entre les mots associés à t_i d'après le dictionnaire. Dans le cas contraire, $C(t_i)$ devenant différent de 0, les probabilités sont évaluées à l'aide de l'égalité (4.3). Certains couples (w_i, t_i) n'apparaissant pas dans le corpus d'apprentissage, il devient nécessaire d'appliquer des méthodes de lissage (*cf.* section 1.2.1) pour éviter d'avoir des probabilités $P(w_i|t_i)$ égales à 0. Nous avons testé deux types de lissage. Le premier, qualifié d'additif, calcule les probabilités par :

$$P(w_i|t_i) = \frac{C(w_i, t_i) + \delta}{\delta \times |\mathcal{V}_{t_i}| + C(t_i)} \tag{4.5}$$

où δ est une constante fixée à 0,5 ou 1, et $|\mathcal{V}_{t_i}|$ représente le nombre de mots pouvant être étiquetés par t_i selon le vocabulaire de l'étiqueteur. Le deuxième, dit de Good-Turing, procède de la manière suivante :

$$P(w_i|t_i) = \frac{C^*(w_i, t_i)}{C(t_i)} \tag{4.6}$$

où $C^*(w_i, t_i)$ est calculé au moyen de l'égalité (1.13) page 20. Dans la pratique, nous avons utilisé une variante, la méthode simple de Good-Turing [Gale et Sampson, 1995], qui tient compte du fait que pour certaines valeurs c, le nombre N_c de couples qui ont été observés c fois dans le corpus peut être nul.

Afin de connaître les paramètres les plus satisfaisants pour lisser les probabilités lexicales, nous avons mesuré le nombre d'étiquettes correctes attribuées par l'étiqueteur sur un corpus de développement de 6 780 mots. Nos résultats montrent que le lissage additif avec $\delta = 1$ est légèrement plus performant (*cf.* annexe B), ce qui contredit une des conclusions présentées dans [Nivre, 2000]. Les différences observées peuvent toutefois être liées aux différences existant entre les corpus étudiés puisque notre corpus de développement contient uniquement 46 mots absents du lexique de notre étiqueteur, ce qui représente 0,7 % du total, alors que celui utilisé dans [Nivre, 2000] est en langue suédoise et contient une proportion de mots inconnus nettement plus élevée de 7,3 %.

4.3.3.2 Probabilités contextuelles

Pour évaluer les probabilités contextuelles, nous avons construit des modèles de langage sur les étiquettes au moyen de la boîte à outils SRILM[6] [Stolcke, 2002]. Le calcul de ces probabilités, contrairement à celui des probabilités lexicales, concerne des N-grammes, ce qui autorise l'usage du repli pour lisser les probabilités :

$$P_{\text{repli}}(t_i|t_{i-N+1}^{i-1}) = \begin{cases} P^*(t_i|t_{i-N+1}^{i-1}) & \text{si } C(t_{i-N+1}^{i}) > k \\ \alpha(t_{i-N+1}^{i-1}) \times P_{\text{repli}}(t_i|t_{i-N+2}^{i-1}) & \text{sinon} \end{cases} \tag{4.7}$$

Lorsque le nombre d'occurrences de t_{i-N+1}^{i} dans le corpus d'apprentissage a dépassé un seuil critique k, fixé à 0 par exemple, les probabilités contextuelles $P_{\text{repli}}(t_i|t_{i-N+1}^{i-1})$ sont

[5] Cela peut se produire pour des étiquettes rares, notamment celles correspondant au subjonctif imparfait.

[6] Version 1.5.0 distribuée par SRI International et disponible gratuitement à des fins de recherche.

estimées à partir des nombres d'occurrences constatées dans le corpus d'apprentissage. De manière à conserver une masse de probabilité pour les séquences peu ou même non observées, les probabilités $P^*(t_i|t_{i-N+1}^{i-1})$ incluent un facteur d'escompte :

$$P^*(t_i|t_{i-N+1}^{i-1}) = \frac{C^*(t_{i-N+1}^{i})}{C(t_{i-N+1}^{i-1})} \ . \tag{4.8}$$

Nous avons testé pour notre part les méthodes disponibles dans SRILM : les lissages de Katz, Witten-Bell, Kneser-Ney et Kneser-Ney modifié, dont la description peut être trouvée dans [Chen et Goodman, 1998], et celui de Ristad [Ristad, 1995].

Pour les séquences observées moins de k fois dans le corpus, les probabilités sont estimées à partir des N-grammes d'ordre inférieur. $P_{\text{repli}}(t_i|t_{i-N+2}^{i-1})$ est alors pondérée par un coefficient $\alpha(t_{i-N+1}^{i-1})$ afin que les estimations des probabilités conditionnelles respectent les contraintes stochastiques, *i.e.*,

$$\sum_{t_i} P_{\text{repli}}(t_i|t_{i-N+1}^{i-1}) = 1 \ . \tag{4.9}$$

Ce coefficient est déterminé à partir de la masse de probabilité réservée aux N-grammes peu observés en remarquant que (4.9) est équivalent à :

$$\sum_{t_i \text{ t.q. } C(t_{i-N+1}^{i})>k} P^*(t_i|t_{i-N+1}^{i-1}) + \sum_{t_i \text{ t.q. } C(t_{i-N+1}^{i})\leq k} \alpha(t_{i-N+1}^{i-1}) \times P_{\text{repli}}(t_i|t_{i-N+2}^{i-1}) = 1 \ , \tag{4.10}$$

ce qui donne :

$$\alpha(t_{i-N+1}^{i-1}) = \frac{1 - \sum_{t_i \text{ t.q. } C(t_{i-N+1}^{i})>k} P^*(t_i|t_{i-N+1}^{i-1})}{\sum_{t_i \text{ t.q. } C(t_{i-N+1}^{i})\leq k} P_{\text{repli}}(t_i|t_{i-N+2}^{i-1})} \ . \tag{4.11}$$

Pour calculer $P_{\text{repli}}(t_i|t_{i-N+1}^{i-1})$, SRILM débute par les probabilités unigrammes $P_{\text{repli}}(t_i)$ en les assimilant à $P^*(t_i)$, puis calcule les probabilités contextuelles utilisant un ordre de plus en plus grand.

Dans le but d'optimiser les calculs de l'étiqueteur, nous avons mesuré le nombre d'étiquettes correctes sur le corpus de développement déjà employé pour paramétrer les probabilités lexicales (*cf.* annexe B). En testant différents ordres N et en comparant plusieurs méthodes de lissage, nos expériences montrent qu'un ordre trigrammes suffit et qu'il y a peu de différences observées entre les différentes techniques de lissage. En n'utilisant aucun lissage, nous n'observons notamment pas de dégradation significative des performances, ce qui montre donc que, dans le cas de l'étiquetage où le nombre de classes envisagé reste restreint, le lissage est beaucoup moins primordial qu'il ne l'est pour les modèles N-grammes de mots utilisés en RAP. Nous avons finalement adopté un ordre de 3, qui est celui généralement employé pour les étiqueteurs stochastiques [Merialdo, 1994 ; Brants, 2000], et un lissage de Katz pour le calcul des probabilités contextuelles.

4.3.4 Utilisation du modèle pour l'étiquetage

Pour étiqueter un nouveau texte, celui-ci est tout d'abord préédité, de manière à avoir une cohérence vis-à-vis du dictionnaire de l'étiqueteur. Cette opération consiste à découper chaque groupe de souffle en mots, en reconnaissant des locutions intégrées dans le lexique, telles que « *parce que* » ou « *a priori* ». L'étiqueteur est ensuite utilisé sur le texte ainsi transformé, de manière à estimer pour chaque groupe de souffle la séquence d'étiquettes la plus probable $\hat{t}_1^n$ parmi toutes celles possibles t_1^n d'après le dictionnaire. Le choix de $\hat{t}_1^n$ pour chaque suite de mots w_1^n peut être réalisé selon deux critères [Merialdo, 1994]. Le premier, reposant sur l'utilisation de l'algorithme de Viterbi, consiste à choisir globalement la séquence d'étiquettes la plus probable selon un critère MAP :

$$\hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n|w_1^n) \ . \tag{4.12}$$

La seconde méthode, reposant sur l'algorithme *forward-backward* [Rabiner, 1989], choisit séparément pour chaque mot w_i l'étiquette $\hat{t}_i$ qui lui paraît la plus probable :

$$\hat{t}_i = \arg\max_{t_i} P(t_i|w_1^n) \ . \tag{4.13}$$

Cette seconde approche minimise le nombre d'erreurs par mot, tout en requérant un nombre de calculs supplémentaires très réduit par rapport à la première[7]. Nous utilisons donc l'algorithme *forward-backward* durant la phase d'étiquetage, même si l'amélioration du nombre d'étiquettes correctes apportée par cette méthode reste très réduite (de l'ordre de 0,1 % en variation absolue sur nos données).

Après avoir exposé le paramétrage et le mode de fonctionnement de notre étiqueteur, nous mesurons la qualité de ses résultats sur des transcriptions, aussi bien manuelles qu'automatiques.

4.4 Évaluation de l'étiquetage automatique sur des documents oraux

Cette section expose les principes sur lesquels est basée l'évaluation d'un étiqueteur, avant de décrire le protocole adopté pour mesurer les performances de l'outil que nous avons développé. Nous présentons enfin les résultats obtenus sur le corpus oral Ester.

4.4.1 Principes

La mesure quantitative de la qualité d'un étiqueteur consiste à calculer :

$$pcorr = \frac{\text{Nb d'étiquettes correctes}}{\text{Nb total de mots du corpus}} \ . \tag{4.14}$$

[7] La complexité des opérations permettant d'obtenir les probabilités nécessaires est dans les deux cas en $\mathcal{O}(m^2 n)$ où m est le nombre maximum d'états précédant l'état courant et n est la longueur de la séquence à analyser.

Ce critère qui paraît simple à appliquer de prime abord est problématique quant à la définition d'une étiquette correcte. Les jeux d'étiquettes habituellement employés, y compris le nôtre, conviennent pour des applications d'ingénierie mais ne s'appuient pas sur une théorie linguistique précise. Le choix de la catégorie correcte pour certains emplois de mots comme « *tout* », « *beaucoup* », « *que* », ou encore les numériques cardinaux (« *un* », « *deux* », *etc.*) reste ainsi sujet à discussion entre linguistes, ce qui rend délicate la constitution d'un étiquetage de référence. Une solution qui peut être adoptée consiste à être laxiste sur la définition d'une étiquette correcte, en ne rejetant que les erreurs flagrantes. Une difficulté supplémentaire réside dans les différentes segmentations possibles d'une même phrase, certaines suites de mots pouvant ainsi être considérées ou non comme des locutions. La séquence « *en dehors* » peut ainsi être considérée comme une locution adverbiale ou bien être décomposée en une préposition, suivie d'un nom commun. Une solution possible peut consister à avoir un double marquage, en constituants et en locutions [Adda *et al.*, 1999].

Si la définition d'un étiquetage de référence peut être malaisée pour un texte produit par un humain, elle l'est bien plus encore pour une transcription produite par un système de RAP. Pour certains groupes de souffle devenus agrammaticaux par les erreurs de reconnaissance, produire un étiquetage de référence se révèle même impossible. Nous proposons donc deux critères d'évaluation :

$$pcorrTot = \frac{\text{Nb d'étiquettes correctes attribuées aux mots correctement transcrits}}{\text{Nb total de mots de la transcription de référence}} \tag{4.15}$$

$$pcorrRec = \frac{\text{Nb d'étiquettes correctes attribuées aux mots correctement transcrits}}{\text{Nb total de mots correctement transcrits}} . \tag{4.16}$$

Ces deux critères permettent d'utiliser comme étiquetage de référence la transcription manuelle correspondant au document à transcrire. Afin de les calculer, la transcription automatique annotée par l'étiqueteur morpho-syntaxique est alignée avec celle de référence étiquetée manuellement (Fig. 4.3), en utilisant le même algorithme d'alignement que celui qui est employé pour calculer le WER (*cf.* section 1.3.2). En ne prenant en compte dans leur calcul que les mots correctement reconnus, les critères ignorent les mots mal transcrits mais ils permettent de vérifier que l'étiqueteur morpho-syntaxique n'est pas perturbé par les erreurs de reconnaissance pour les mots qui sont correctement transcrits.

4.4.2 Protocole

Afin de tester notre étiqueteur, nous avons constitué un étiquetage de référence pour la transcription manuelle d'une émission d'1 heure de France Inter. Cet extrait du corpus ESTER est constitué de 11 300 mots répartis en 518 groupes de souffle. Nous avons effectué deux segmentations différentes du document : l'une en groupes de souffle, l'autre sous la forme de phrases. L'étiquetage a été réalisé en s'inspirant des critères d'assignation des étiquettes morpho-syntaxiques adoptés pour la campagne d'évaluation GRACE [Lecomte, 1997], même si des choix différents ont été faits puisque

REF:		HYP:	
il	PPER3S	il	PPER3S
fait	VINDP3S	fait	VINDP3S
sortir	VINF	sortir	VINF
le	DETMS	le	DETMS
groupe	NCMS	groupe	NCMS
de	PREP	de	PREP
la	DETFS		
cave	NCFS		
et	COO	quatre	CAR
l'	PPER3S	vingt	CAR
amène	VINDP3S	mai	NCMS
dans	PREP	dans	PREP
un	DETMS	un	DETMS
bâtiment	NCMS	bâtiment	NCMS
au-dessus	ADV		

FIG. 4.3: *Alignement de la transcription automatique étiquetée par l'étiqueteur (*HYP*) avec la transcription de référence étiquetée manuellement (*REF*)*

notre jeu d'étiquette est beaucoup moins fin que celui du projet GRACE. Globalement, nous avons regroupé peu de mots en locutions, dans la mesure où notre étiqueteur en reconnaît peu.

La mesure de la qualité de l'étiquetage a été effectuée sur deux versions de la transcription d'une émission d'information : l'une réalisée manuellement et l'autre produite par notre système de RAP. De manière à évaluer avant tout l'influence des mots mal reconnus, nous avons supprimé les majuscules et les marques de ponctuation de la transcription manuelle, et procédé à une resegmentation en groupes de souffle. Notons que le document audio étudié fait partie du corpus d'apprentissage du système de transcription. Bien que les expériences menées au chapitre suivant soient faites à partir d'autres transcriptions étiquetées, le WER calculé sur cet extrait, égal à 22 % et donc similaire au nombre de mots mal reconnus sur les corpus de développement et de test, nous laisse penser que les performances que nous mesurons ici sont représentatives des données que nous manipulons par la suite.

La détermination des étiquettes correctes, que ce soit pour une transcription manuelle ou automatique, est réalisée en comparant pour chaque mot l'étiquette de référence avec celle trouvée par l'étiqueteur. Seules les étiquettes strictement identiques sont dans un premier temps comptabilisées comme correctes. Les classes proposées par l'étiqueteur peuvent parfois être considérées comme acceptables même si elles ne sont pas exactement identiques aux étiquettes de référence. Nous donnons ainsi pour chaque critère *pcorr*, *pcorrTot* et *pcorrRec* un deuxième chiffre où nous intégrons dans les étiquettes acceptables deux nouveaux cas de figure. Le premier correspond aux confu-

sions entre participe passé et adjectif ; `ADJMS`[8] et `VPARPMS`[9] sont alors considérés comme équivalents mais pas `ADJMS` et `VPARPFS`[10] qui diffèrent selon le genre. Excepté lorsque le participe passé suit l'auxiliaire « *avoir* », la distinction entre ces deux catégories est en effet très délicate. Le second cas correspond aux étiquettes attribuées pour les noms communs et noms propres et qui ne sont pas en contradiction avec les étiquettes de référence. Par exemple, dans le cas où l'étiqueteur choisit le label `NPSIG`[11] qui ne fait pas de distinction sur le genre, nous le considérons comme acceptable si l'étiquette de référence correspond à un nom propre singulier, qu'il soit féminin ou pluriel.

4.4.3 Résultats

L'application de notre étiqueteur sur des transcriptions manuelle et automatique segmentées en groupes de souffle conduit aux résultats présentés dans le tableau 4.1. Si nous nous en tenons aux pourcentages donnés pour les étiquettes acceptables, nos résultats sont comparables aux 95-98 % habituellement retenus pour les documents écrits, ce qui établit clairement que l'étiquetage produit est bon dans l'ensemble que ce soit pour la transcription manuelle — où on ne trouve ni majuscules, ni marques de ponctuation, ni une segmentation linguistique en phrases — ou pour l'automatique, où figurent de surcroît des erreurs de reconnaissance. En examinant les 481 erreurs d'étiquetage produites pour la transcription manuelle en ignorant les erreurs acceptables, on constate que 52 sont dues à des mots qui ne sont pas présents dans le dictionnaire de notre étiqueteur, que 19 sont dues à une confusion entre nom commun et nom propre qui serait probablement corrigée à l'aide de majuscules et que 86 sont expliquées par une mauvaise *tokenisation* de notre étiqueteur. Dans ce dernier groupe d'erreurs, on trouve ainsi des mots composés, comme « *proche-orient* » interprété comme un adjectif suivi d'un nom commun, des entités nommées, telles que « *le monde* » étiquetées par un déterminant suivi par un nom commun, et des locutions comme « *y compris* », reconnu comme étant un adverbe suivi d'un participe passé. Près du tiers des erreurs commises par notre étiqueteur ne sont donc pas dues à l'étiqueteur proprement dit mais à notre lexique et au module de reconnaissance de groupes de mots.

Les jeux d'étiquettes faisant parfois une distinction moins fine que le nôtre au niveau des informations morphologiques, nous avons poursuivi l'analyse des erreurs d'étiquetage en regardant la proportion qui sont dues à des erreurs sur le genre, le nombre ou la conjugaison. Les pourcentages présentés dans les dernières lignes du tableau 4.1 indiquent qu'en considérant des classes simples qui ne portent pas d'informations sur le genre et le nombre, les pourcentages d'étiquettes correctes avoisineraient 97 %. Nous préférons toutefois conserver cette connaissance puisqu'elle se révèle primordiale pour la correction d'erreurs de transcription.

L'étiquetage étant satisfaisant, nous avons voulu savoir si une segmentation en groupes de souffle perturbait notre étiqueteur. Nous avons donc mené des expériences

[8] Adjectif masculin singulier
[9] Verbe au participe passé sous sa forme masculin singulier.
[10] Verbe au participe passé sous sa forme féminin singulier.
[11] Nom propre singulier dont le genre n'est pas précisé.

transcription	manuelle	automatique	
	pcorr	*pcorrTot*	*pcorrRec*
étiquettes correctes	94,3	74,5	94,2
étiquettes acceptables	95,7	75,6	95,7
erreurs ignorées sur le nombre	96,1	76,0	96,1
erreurs ignorées sur le genre	96,3	76,2	96,4
erreurs de conjugaison ignorées	96,1	75,9	96,0
uniquement les erreurs sur la classe grammaticale	96,7	76,5	96,8

TAB. 4.1: *Évaluation de notre étiqueteur (en pourcentages)*

transcription	manuelle	automatique	
	pcorr	*pcorrTot*	*pcorrRec*
étiquettes correctes	89,0	70,4	89,0
étiquettes acceptables	90,6	71,7	90,7

TAB. 4.2: *Évaluation de l'approche naïve (en pourcentages)*

similaires sur la transcription manuelle segmentée en phrases. Nous avons alors constaté que les pourcentages d'étiquettes correctes étaient équivalents quelle que soit la segmentation, ce qui montre que ce n'est pas un facteur perturbant pour l'étiquetage.

Afin de mesurer la complexité de la tâche d'étiquetage sur notre corpus, nous avons utilisé une méthode naïve utilisant le même dictionnaire que notre étiqueteur et qui attribue à chaque mot l'étiquette qui lui est la plus souvent associée dans le corpus d'apprentissage. Nous avons tout d'abord constaté que $71,0\,\%$ des mots du lexique étaient associés à une seule étiquette d'après le dictionnaire. En recourant à cette approche naïve pour étiqueter l'émission étudiée, les résultats (Tab. 4.2) montrent ainsi qu'une bonne part des mots sont déjà associées correctement à leur classe catégorielle et que l'utilisation du contexte trigrammes permet de rectifier la moitié des erreurs commises.

Nous avons enfin mené une dernière série d'expériences en comparant notre étiqueteur au logiciel CORDIAL, généralement considéré comme celui de référence pour le français et qui a déjà donné de bons résultats sur un corpus de parole [Valli et Véronis, 1999]. La comparaison avec cet étiqueteur a nécessité d'établir des conversions d'étiquettes car bien que nous nous sommes inspiré du jeu d'étiquettes de CORDIAL, le nôtre présente quelques différences. Nous avons ainsi fait le choix de préciser systématiquement le genre et le nombre pour les adjectifs et les noms communs alors que CORDIAL se contente par exemple d'une seule étiquette `ADJSIG`[12] pour des adjectifs comme « *optimiste* » qui ne varient pas quand on les accorde suivant le genre. Le jeu de CORDIAL est en revanche plus précis pour certaines classes d'étiquettes puisque nous

[12] Adjectif singulier invariant en genre.

transcription		manuelle	automatique	
		pcorr	*pcorrTot*	*pcorrRec*
notre étiqueteur	étiquettes acceptables	95,7	75,6	95,7
	confusions entre NC et NP ignorées	95,9	75,8	95,9
CORDIAL	étiquettes acceptables	90,7	71,7	90,6
	confusions entre NC et NP ignorées	95,0	75,3	95,2

TAB. 4.3: *Comparaison de notre étiqueteur avec* CORDIAL *(en pourcentages)*

avons regroupé les adjectifs interrogatifs et les adjectifs indéfinis parmi les déterminants. Le tableau 4.3 présente les résultats obtenus par les deux étiqueteurs sur le corpus de test. Il établit que notre étiqueteur a des résultats supérieurs à CORDIAL en ce qui concerne les pourcentages d'étiquettes acceptables. Cette conclusion doit toutefois être prise avec précaution. D'une part, l'émission que nous avons considérée comme corpus de test commente abondamment la guerre en Irak en 2003, dont il est question dans près de la moitié des documents du corpus d'apprentissage de notre étiqueteur, contrairement à celui de CORDIAL. D'autre part, ces différences s'expliquent par la nature particulière des textes à analyser, pour laquelle CORDIAL n'a pas été spécifiquement conçu. Les manques de ponctuation et surtout de majuscules — ces dernières étant employées comme indice pour détecter des noms propres — sont particulièrement problématiques. En ignorant les confusions entre nom commun et nom propre, les expériences montrent d'ailleurs que les deux étiqueteurs ont des performances comparables.

Cette série d'expérimentations établit que notre étiqueteur est capable de produire une annotation fiable sur des documents transcrits automatiquement, ce qui n'était encore qu'une hypothèse auparavant. Il n'est notamment pas perturbé par les mots mal reconnus pour étiqueter convenablement les autres mots puisque les performances sont similaires entre une transcription sans erreur et une avec près d'un mot sur quatre erroné. Cette propriété est expliquée par le fait que les étiquettes sont attribuées en exploitant des informations de manière très locale. La section suivante présente le comportement de notre étiqueteur pour d'autres jeux d'étiquettes que nous avons employés dans les expériences menées au chapitre 5.

4.4.4 Autres jeux d'étiquettes

Notre étiqueteur vise à introduire des informations explicites sur le genre et le nombre, de manière à corriger principalement les erreurs d'accord présentes dans la transcription. Nous avons, à cette fin, envisagé quatre jeux d'étiquettes supplémentaires (*cf.* annexe A) à celui de 93 étiquettes que nous avons utilisé jusqu'à présent et que nous notons par la suite *norm*. Nous espérons avoir de meilleures performances à l'étiquetage et surtout observer un meilleur comportement pour améliorer la transcription. Pour ne pas avoir à réétiqueter manuellement le corpus d'apprentissage de notre étiqueteur, nous avons choisi ces jeux de manière à ce que chaque classe parmi celles de

transcription	manuelle	automatique	
	pcorr	*pcorrTot*	*pcorrRec*
*red*1	96,4	76,3	96,4
*red*2	95,8	75,7	95,7
*ext*1	96,8	76,7	97,0
*ext*2	96,9	76,8	97,1

TAB. 4.4: *Pourcentages d'étiquettes acceptables pour différents jeux d'étiquettes*

norm corresponde au plus pour un mot donné à une seule étiquette du nouvel ensemble. Cela nous permet ainsi de ne pas avoir à désambiguïser à la main les étiquettes.

Le premier ensemble *red*1 réduit considérablement le nombre d'étiquettes différentes en ne distinguant les mots lexicaux que selon leur genre et leur nombre. L'objectif consiste à avoir un nombre minimum de classes à envisager, tout en espérant avoir suffisamment d'informations pour corriger les erreurs d'accord. Le deuxième ensemble *red*2 introduit des étiquettes particulières pour les auxiliaires « *être* » et « *avoir* » et ne fait plus de distinction sur le mode et le temps, celle-ci étant souvent difficile. Le troisième ensemble *ext*1 étend *norm* par des étiquettes spécifiques pour les mots grammaticaux les plus courants dans notre corpus d'apprentissage. Ceci permet de limiter les risques d'erreurs d'étiquetage pour les mots ambigus et de faire intervenir explicitement dans le calcul des probabilités les mots structurant un groupe de souffle. Un dernier ensemble *ext*2 vise à combiner les avantages de *red*2 et *ext*1 en ajoutant à *red*2 des étiquettes particulières pour les mots grammaticaux les plus courants. Le tableau 4.4 présente le comportement de notre étiqueteur pour chacun des jeux d'étiquettes sur le corpus de test. Le jeu d'étiquettes *red*1 étant peu informatif pour l'attribution des étiquettes en fonction du contexte, nous avons préféré dans un premier temps étiqueter le corpus de test par *norm* avant de convertir les étiquettes obtenues vers l'ensemble *red*1. Le nombre d'étiquettes acceptables est supérieur pour chacun des 4 jeux d'étiquettes aux résultats obtenus lors de nos tests précédents. Dans le cas de *red*1 et *red*2, cette constatation est liée à la réduction des ambiguïtés du fait de la diminution du nombre d'étiquettes ; dans le cas de *ext*1 et *ext*2, elle est expliquée par la suppression des ambiguïtés pour des mots très fréquents. Cette amélioration de la qualité de l'étiquetage sur le corpus de test montre ainsi que nous pouvons obtenir des informations plus fiables encore pour ces quatre jeux d'étiquettes.

4.5 Bilan

En examinant les principales erreurs produites par un système de reconnaissance, nous avons montré d'une part l'intérêt des POS pour rectifier certaines d'entre elles, notamment celles concernant les accords sur le genre et le nombre. Sur l'extrait de transcription automatique que nous avons analysé, nous avons constaté qu'en corrigeant les erreurs d'accord expliquées par une mauvaise prise en compte des dépen-

dances au sein d'un même groupe de souffle nous pouvons espérer une baisse absolue du WER de 1,1 %. D'autre part, nous avons prouvé quantitativement que les étiqueteurs morpho-syntaxiques pouvaient réellement être employés sur des corpus oraux transcrits manuellement ou automatiquement. Nous avons construit un étiqueteur stochastique basé sur les MMC et exploitant un contexte trigrammes qui nous permet d'avoir des résultats comparables à ceux obtenus sur des documents écrits. Cet étiqueteur a montré sa robustesse aux erreurs de transcription, malgré sa simplicité. En produisant des résultats comparables avec les outils standard, il nous donne toute la liberté nécessaire pour l'utiliser par la suite. Le chapitre suivant présente notre méthode et nos expérimentations pour employer au sein d'un système de RAP les informations linguistiques automatiquement produites par notre étiqueteur.

Chapitre 5

Traitement de listes d'hypothèses pour améliorer la transcription

Ce chapitre poursuit l'étude des informations morpho-syntaxiques en RAP que nous avons débutée au précédent. Après avoir montré que les POS constituent des informations *a priori* pertinentes pour corriger une part significative des erreurs de transcription et qu'elles pouvaient être obtenues de manière automatique et fiable sur des hypothèses de transcription, nous les utilisons désormais en post-traitement d'un système de RAP. Notre objectif est ici de montrer expérimentalement que nous pouvons améliorer la qualité de la transcription de manière significative au moyen des POS. Notre approche se distingue de celles précédemment envisagées pour introduire cette connaissance pour le décodage de la parole sur deux principaux aspects. D'une part, nous employons un étiqueteur pour associer aux hypothèses de mots produites leur classe catégorielle la plus probable, en privilégiant ensuite les sorties associées aux successions de POS les plus probables. Les méthodes usuelles, recourant aux modèles N-classes, ne procèdent pas à cette étape de désambiguïsation et envisagent toutes les classes possibles lors de l'estimation des probabilités. D'autre part, nous proposons un nouveau mode de calcul du score attribué à chaque hypothèse de transcription en incluant les POS en post-traitement du système de RAP Irene ; nous montrons expérimentalement son intérêt aussi bien pour diminuer le WER que pour détecter les mots mal reconnus.

Nous présentons dans un premier temps notre méthode pour intégrer les POS dans le score employé par le système de transcription pour sélectionner les hypothèses de succession de mots les plus probables. Nous montrons ensuite à travers différentes expériences que l'utilisation de cette technique en post-traitement permet une amélioration significative de la qualité de la sortie d'un système de RAP, suivant plusieurs critères de codage. Nous exposons dans la section suivante une approche complémentaire qui se concentre sur la discrimination des homophones ne se distinguant que par la flexion

selon le genre et le nombre. Nous indiquons enfin comment les POS aident à la détection des erreurs de transcription.

5.1 Score intégrant les parties du discours

Tout au long de cette étude sur l'apport des informations morpho-syntaxiques pour la transcription de la parole, nous envisageons les POS en post-traitement d'un système de RAP en opérant sur des listes de $\mathcal{N}$ meilleures hypothèses (*cf.* sections 5.2 et 5.4) ou sur des ensembles d'hypothèses homophones (*cf.* section 5.3). Bien que ces sorties soient moins informatives que des structures de données plus complexes telles que des graphes de mots, chaque hypothèse de ce type de liste peut être vue comme du texte standard, ce qui permet d'appliquer facilement un étiqueteur morpho-syntaxique. En outre, opérer sur des graphes de mots demanderait une augmentation importante du temps de calcul ; dans le cas de notre étiqueteur, le calcul des probabilités contextuelles $P(t_i|t_{i-2}t_{i-1})$ requiert ainsi l'extension de chaque nœud associé à un trigramme de mots en autant de nœuds qu'il existe de POS possibles pour cette suite de mots.

Pour combiner les informations morpho-syntaxiques avec celles du MA et du ML de mots, deux solutions peuvent être envisagées. Une première consiste à intégrer les POS au niveau du ML, autrement dit au niveau du mot ; cette approche, en recourant notamment aux modèles N-classes, a donné lieu à des résultats peu probants. Une seconde que nous proposons consiste à utiliser les POS comme une connaissance à part entière en les combinant avec celles du MA et du ML au niveau du groupe de souffle. Elle présente notamment l'avantage par rapport aux méthodes classiques de pouvoir segmenter différemment les séquences de mots et celles de POS, et conduit expérimentalement à une amélioration plus significative de la transcription, comme nous allons le montrer ici.

Dans un premier temps, les hypothèses w_1^n proposées par le système de RAP pour chaque groupe de souffle sont étiquetées par notre étiqueteur morpho-syntaxique par leur séquence la plus probable t_1^m. Nous attribuons alors un score morpho-syntaxique à w_1^n :

$$\log P(t_1^m) = \sum_{i=1}^{m} \log P(t_i|t_{i-N+1}^{i-1}) \ . \tag{5.1}$$

Il est à noter que le nombre d'étiquettes m peut différer de celui des mots n. En effet, notre étiqueteur reconnaît des suites de mots telles que « *parce que* » comme des locutions. De plus, afin de favoriser la prise en compte des relations à longue distance par $P(t_i|t_{i-N+1}^{i-1})$, nous avons regroupé les mots étiquetés consécutivement comme étant des cardinaux (`CAR`) par une seule étiquette `CAR`. De même, des mots consécutifs reconnus comme étant des noms propres sont associés à une étiquette correspondant à un nom propre unique. Nous avons enfin fait le choix d'ignorer dans le calcul du score les mots reconnus comme des interjections, tels que « *bon* », « *ben* » et « *euh* » en constatant lors d'expériences préliminaires que les ignorer conduisait à une légère diminution du WER. Ceci est justifié par le fait que les interjections sont une classe grammaticale à part, dont l'emploi présente peu de connexions avec le restant de la phrase.

Ce score morpho-syntaxique est intégré dans le critère de sélection des meilleures hypothèses par le système de RAP (*cf.* équation (1.11) page 16) en lui associant une pondération appropriée. Le score combiné pour un groupe de souffle w_1^n correspondant au vecteur acoustique A devient alors :

$$s(w_1^n) = \log P(w_1^n|A) + \alpha \log P(w_1^n) + \beta \log P(t_1^m) + \gamma n \tag{5.2}$$

où α est le facteur de pondération du ML, β celui du score morpho-syntaxique et γ la pénalité d'insertion de mot. La prise en compte des informations morpho-syntaxiques permet ainsi de pénaliser plus explicitement des séquences de mots peu probables tels qu'un nom commun au pluriel suivant un adjectif au singulier.

Suivant la manière dont est calculée la probabilité $P(w_1^n)$ dans notre système de transcription, nous construisons un ML N-grammes à base de POS pour déterminer le score morpho-syntaxique. Nous avons considéré un ordre $N = 7$, supérieur au contexte quadrigrammes du ML utilisé dans les dernières passes du système IRENE, de façon à obtenir un historique de taille plus importante que celui du ML de mots et ainsi pouvoir prendre en compte des relations à plus longue distance. Cette extension du contexte est autorisée par le nombre très réduit de POS possibles par rapport à la taille d'un vocabulaire d'un système de RAP. Nous avons en outre conservé par défaut la méthode de lissage employée lors du calcul des probabilités pour l'étiquetage, *i.e.*, celle de Kneser-Ney. Des expériences menées en réordonnant les listes des 100 meilleures hypothèses avec $s(w_1^n)$, selon un critère MAP (*cf.* section 5.2.1), montrent ainsi une influence limitée de l'ordre et du lissage. En testant avec 4 types de lissage différents (Katz, Ristad, Kneser-Ney en version modifiée ou non) et les 3 ordres $N = 5$, $N = 7$ et $N = 9$, nous avons ainsi constaté une variation en absolu du WER de 0,1 point sur le corpus de développement.

Tout au long des sections suivantes, nous utilisons le score $s(w_1^n)$ afin de réordonner des listes d'hypothèses et de détecter des erreurs de transcription.

5.2 Réordonnancement des $\mathcal{N}$ meilleures hypothèses

Suivant le principe exposé à la section précédente, nous prenons en compte les informations morpho-syntaxiques dans le score global après avoir étiqueté selon les POS les listes des 100 meilleures hypothèses produites par le système IRENE. Ces listes sont informatives puisqu'en comparant les scores oracles[1] obtenus par différents types de sortie, nous constatons que peu d'informations sont perdues en se limitant à des listes des 100 meilleures hypothèses (Tab. 5.1) et que l'on peut espérer un gain important du WER en réordonnant ces listes par rapport à la transcription choisie initialement par le système IRENE. Ainsi, sur le corpus de développement, si cette dernière possède un WER de 21,6 %, il est possible d'atteindre un taux d'erreur égal à 14,2 % (resp. 13,9 %) en choisissant systématiquement la meilleure hypothèse pour chaque groupe de souffle

[1]Scores les meilleurs que l'on puisse obtenir à partir d'une représentation de données. Dans notre cas, il s'agit des WER minimaux.

	1-best	100-best	graphe de mots
développement	21,6	14,2	13,9
test	19,9	12,4	11,9

TAB. 5.1: *WER oracles (en pourcentages) calculés à partir de la transcription proposée par le système* IRENE*, la liste des 100 meilleures hypothèses et le graphe de mots de la dernière passe*

parmi les listes de 100 hypothèses (resp. les graphes de mots) proposées en sortie du système de RAP.

Les 100 meilleures hypothèses trouvées par le système de RAP pour chaque groupe de souffle sont annotées par notre étiqueteur morpho-syntaxique puis associées au score défini par (5.2) qui inclut toutes les sources disponibles de connaissances. Ces listes sont réordonnées selon trois critères différents (Fig. 5.1) : le critère MAP, la minimisation du nombres d'erreurs de transcription [Stolcke *et al.*, 1997] et le décodage par consensus [Mangu *et al.*, 2000]. Ces deux derniers types de décodage souvent utilisés dans les systèmes de RAP actuels visent à réduire le WER mais augmentent généralement le SER.

Au cours de nos diverses expérimentations sur le réordonnancement de listes d'hypothèses, nous employons deux corpus extraits de la partie du corpus ESTER dédiée au développement (Tab. 3.2 page 55). Le premier, utilisé comme corpus de développement, permet de déterminer empiriquement les paramètres α, β et γ qui minimisent le WER pour l'ensemble des émissions qui le composent. Le choix de ces coefficients n'est pas difficile à effectuer car nous avons constaté au cours de nos expériences qu'un minimum global apparaissait rapidement en faisant varier les trois paramètres simultanément. La figure 5.2 montre ainsi l'évolution du WER sur le corpus de développement en fonction de α et β, en choisissant à chaque fois le coefficient γ pour lequel on observe le WER minimal. On observe ainsi sur ce graphique que le minimum global est atteint quand $\beta > 0$, *i.e.*, quand on prend en considération les POS. Une fois les trois paramètres optimisés, ils sont fixés et testés pour chaque méthode sur le deuxième ensemble d'émissions qui fait office de corpus de test. Le modèle calculant le score morpho-syntaxique $\log P(t_1^m)$ est quant à lui établi à partir du même corpus d'apprentissage que celui qui a permis la construction de notre étiqueteur (*cf.* section 4.3.2).

Cette section décrit tout d'abord l'évolution du WER en fonction des trois différents critères que nous avons testés. Elle établit ensuite que le recours aux informations morpho-syntaxiques permet d'obtenir des transcriptions plus facilement exploitables par des méthodes de TAL. Nous montrons enfin à l'aide d'une expérience menée sur la reconnaissance de caractères manuscrits de textes écrits en anglais que nos résultats sont applicables pour d'autres langues que le français.

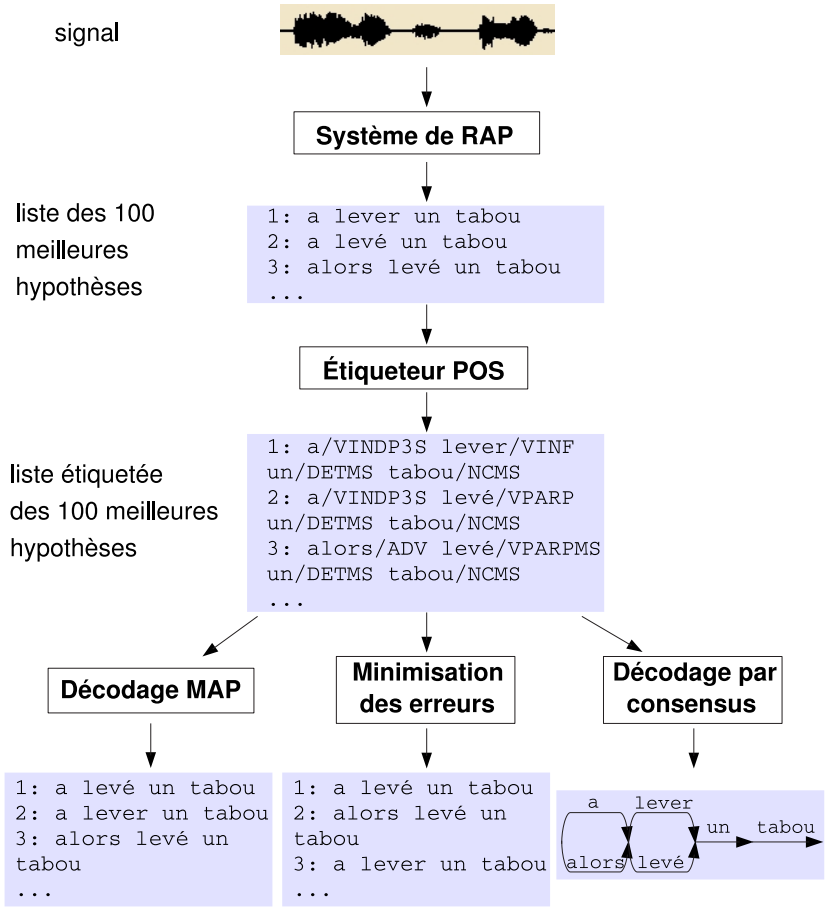

FIG. 5.1: *Principe du réordonnancement des 100 meilleures hypothèses*

5.2.1 Décodage MAP

Le critère MAP consiste à rechercher parmi les $\mathcal{N}$ hypothèses proposées $W_{1,\mathcal{N}}$ pour un groupe de souffle donné celle $\hat{W}$ qui vérifie[2] :

$$\hat{W} = \arg \max_{W \in W_{1,\mathcal{N}}} score_{\text{glob}}(W) \ . \tag{5.3}$$

Pour tester la validité de ce critère et de notre score incluant les POS, nous avons mené plusieurs expérimentations, en confrontant nos résultats avec les méthodes usuelles.

5.2.1.1 Premiers résultats

Nous comparons dans un premier temps le WER obtenu avec notre système de RAP selon que l'on inclut on non le score morpho-syntaxique $\beta \log P(t_1^m)$. En optimisant les

[2]Si $score_{\text{glob}}$ est défini de telle manière à ce que sa valeur est d'autant plus grande que l'hypothèse est correcte.

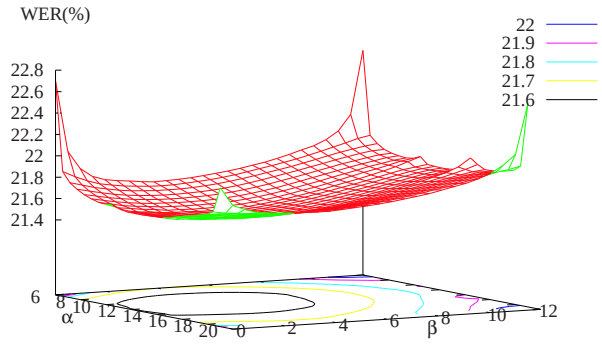

FIG. 5.2: *Évolution du WER en fonction des paramètres α et β avec le jeu d'étiquettes $ext2$ et sur le corpus de développement*

étiquettes	Sub	Del	Ins	WER	α	β	γ
sans POS	13,5	4,5	2,0	19,9	12	0	0
$norm$	13,0	4,1	2,2	19,3	12	4	5
$red1$	13,1	4,2	2,2	19,5	13	3	5
$red2$	13,0	4,1	2,2	19,2	12	4	5
$ext1$	12,9	4,1	2,2	19,2	11	4	5
$ext2$	12,9	3,9	2,3	19,1	10	4	5

TAB. 5.2: *Résultats obtenus sur le corpus de test pour différents jeux d'étiquettes et avec un ordre $N = 7$. Sub, Del et Ins représentent respectivement les taux d'erreur de substitution, de suppression et d'insertion*

coefficients α, β et γ pour chacun des jeux d'étiquettes sur le corpus de développement, nous constatons une baisse du WER sur le corpus de test pour tous nos jeux d'étiquettes[3] (Tab. 5.2). Cette évolution du taux d'erreur se révèle néanmoins assez sensible aux étiquettes utilisées. Pour le meilleur ensemble, *i.e.*, $ext2$, nous observons une baisse absolue du WER de 0,8 % ; c'est dorénavant ce jeu d'étiquettes que nous utilisons dans toutes les expériences qui suivent sur le réordonnancement d'hypothèses.

5.2.1.2 Comparaison avec les ML N-classes

Afin de montrer l'efficacité du score que nous proposons, nous comparons les résultats que nous avons obtenus avec ceux donnés par la méthode qui est généralement employée pour prendre en compte les POS, *i.e.*, les ML N-classes. Ces ML sont généra-

[3] Définis en section 4.4.4 et dans l'annexe A.

système de RAP initial	$s(w_1^n)$	ML N-classes
19,9	19,1	19,5

TAB. 5.3: *WER(%) mesurés sur les données de test*

lement introduits au cours du processus de transcription au moyen d'une interpolation linéaire avec un ML de mots [Maltese et Mancini, 1992]. La recherche de la meilleure hypothèse se fait à l'aide de la formulation classique de la RAP (*cf.* équation (1.11) page 16) en calculant la probabilité de la séquence de mots de la manière suivante :

$$P(w_1^n) = \prod_{i=1}^{n} [\lambda P_{\text{mots}}(w_i|w_1^{n-1}) + (1-\lambda) P_{\text{POS}}(w_i|w_1^{n-1})] \tag{5.4}$$

avec :

$$P_{\text{POS}}(w_i|w_1^{n-1}) = \sum_{t_1 \in \mathcal{T}_1 \ldots t_i \in \mathcal{T}_i} [P(w_i|t_i) P(t_i|t_{i-N+1}^{i-1})] \ . \tag{5.5}$$

λ est un coefficient déterminant le poids du ML M-grammes par rapport au ML N-classes et $\mathcal{T}_i$ représente l'ensemble des étiquettes possibles pour le mot w_i, suivant le dictionnaire de notre étiqueteur. Nous utilisons au cours de nos expériences un ML de mots d'ordre $M = 4$ et un ML de classes d'ordre $N = 7$.

La variation des paramètres α, γ et λ sur le corpus de développement montre une baisse du WER quand on prend en compte le modèle N-classes. En fixant λ à 0,7, nous obtenons sur le corpus de test une baisse absolue du WER de 0.4 % (Tab. 5.3, dernière colonne), ce qui reste une diminution deux fois inférieure à celle observée précédemment (Tab. 5.3, deuxième colonne) avec le score $s(w_1^n)$ établi par l'égalité (5.2).

Une différence importante des deux approches testées est l'utilisation des probabilités lexicales $P(w_i|t_i)$ par les ML N-classes. Pour étudier l'influence de ces valeurs, deux nouveaux scores combinant les informations du MA, du ML et celles apportées par les POS sont définis. Le premier, calculé par :

$$s'(w_1^n) = \log P(w_1^n|A) + \alpha \log P(w_1^n) + \beta[\sum_{i=1}^{n} \log P(w_i|t_i) + \log P(t_1^m)] + \gamma n \ , \tag{5.6}$$

prend en compte les probabilités lexicales après désambiguïsation[4]. Le second considère toutes les séquences possibles plutôt que la meilleure et est déterminé par :

$$s''(w_1^n) = \log P(w_1^n|A) + \alpha \log P(w_1^n) + \beta[\sum_{i=1}^{n} \log P_{\text{POS}}(w_i|w_1^{i-1}) + \log P(t_1^m)] + \gamma n \tag{5.7}$$

où P_{POS} est obtenue par (5.5). Les résultats obtenus avec ces deux scores (Tab. 5.4, lignes 3 et 4) montrent une légère amélioration du WER en prenant en compte les probabilités

[4]Le nombre de $P(w_i|t_i)$ à additionner peut en réalité être différent de n dans le cas où l'étiqueteur reconnaît des locutions dans w_1^n.

système de RAP initial	19,9
probabilités contextuelles + désambiguïsation ($s(w_1^n)$)	19,1
probabilités lexicales et contextuelles + désambiguïsation ($s'(w_1^n)$)	19,0
probabilités lexicales et contextuelles ($s''(w_1^n)$)	19,0

TAB. 5.4: *WER(%) mesurés sur les données de test*

contextuelles. Ils ne montrent pas clairement l'intérêt d'effectuer une désambiguïsation, ce qui peut être lié au fait que généralement la meilleure suite de POS proposée par l'étiqueteur possède une probabilité beaucoup plus grande que celle des autres. En revanche, ils établissent clairement que l'interpolation linéaire des log-probabilités est plus efficace que celle des probabilités.

5.2.1.3 Tests de significativité

Nos expériences précédentes ont établi une baisse globale du WER quand on utilise la connaissance des POS. En examinant individuellement chaque émission du corpus de test, nous avons constaté qu'une diminution est toujours observée, à l'exception d'un programme diffusé sur RTM et d'une durée de 20 minutes. Bien que nos données représentent un volume conséquent, il convient cependant de recourir à des tests statistiques pour s'assurer que cette amélioration des performances ne soit pas liée au hasard.

Afin de comparer le système de RAP sans les POS avec celui les incluant, les tests statistiques confrontent le nombre d'erreurs de transcription produites par ces deux systèmes sur plusieurs régions [Gillick et Cox, 1989]. Ces tests nécessitant que les erreurs faites sur des segments différents soient indépendantes, l'unité choisie est généralement le groupe de souffle puisque les MA et le ML traitent chaque groupe de souffle de manière isolée. Les tests étudient le comportement de la variable aléatoire $Z_i = E^i_{\text{sans POS}} - E^i_{\text{avec POS}}$ représentant la différence entre les nombres d'erreurs observées E_i entre les deux systèmes à comparer sur le segment i. Ils évaluent la probabilité de l'hypothèse H_0, stipulant que les différences observées sont dues au hasard. Nous souhaitons quant à nous établir que l'amélioration est bien significative, ce qui conduit à nous intéresser plutôt à la probabilité que H_0 soit rejetée.

Le premier test que nous employons est le *t-test* pour des séries appariées, celui généralement utilisé dans le cadre de la RAP (voir par exemple [Kim, 2004]). Il suppose que Z suit une loi normale centrée sur 0 et nécessite le calcul de :

$$t = \frac{\hat{\mu}_Z}{\hat{\sigma}_Z/\sqrt{n}} \tag{5.8}$$

où $\hat{\mu}_Z$ représente la moyenne empirique calculée sur l'ensemble de test constitué de n segments :

$$\hat{\mu}_Z = \sum_{i=1}^{n} \frac{Z_i}{n} \tag{5.9}$$

et $\hat{\sigma}_Z$ est l'écart-type empirique établi sur l'ensemble de test :

$$\hat{\sigma}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \hat{\mu}_Z)^2 \ . \tag{5.10}$$

Il est montré dans le cas général que si Z suit une loi normale centrée sur 0, la variable aléatoire t suit une distribution de Student avec $n-1$ degrés de liberté. Le test de rejet de H_0 avec un niveau de confiance α[5] revient alors à déterminer $P(|T| \geq |t|) \leq \alpha$ où T suit la loi de Student et t est estimé sur notre corpus de test.

Le deuxième test que nous utilisons est le *test de Wilcoxon* pour des séries appariées qui présente l'avantage par rapport au premier de faire une hypothèse moins forte sur Z puisqu'il suppose simplement que Z suit une distribution continue, symétrique et centrée sur 0 sous l'hypothèse H_0. Il est souvent préféré pour cette raison au *paired t-test* dans des domaines comme la recherche d'information [Hull, 1993] et est également employé en RAP, quoique de manière plus marginale (voir par exemple [Heeman et Allen, 1999]). Ce test implique le calcul de :

$$t = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}} \tag{5.11}$$

avec :

$$R_i = sign(Z_i) \times rank(|Z_i|) \tag{5.12}$$

où *sign* représente la fonction signe[6] et *rank* est la fonction donnant le classement de $|Z_i|$ dans l'ordre décroissant pour chaque segment i du corpus de test. Il est montré que sous les hypothèses du test, t suit une loi normale standard. Le test d'acceptation de H_0 avec un niveau de confiance α se ramène alors au calcul de $P(|T| \geq |t|) \leq \alpha$, où T suit une loi normale standard et t est calculé sur notre corpus de test.

L'examen de la variable aléatoire Z_i sur les données que nous étudions suggère qu'elle suive une loi normale avec une variance faible (Fig. 5.3), ce qui nous permet d'utiliser les deux méthodes de test. Nous comparons ainsi deux à deux les WER obtenus avec des systèmes calculant différemment le score des hypothèses. Les résultats montrent que pour les deux tests l'hypothèse H_0 est rejetée avec un indice de confiance $\alpha \leq 10^{-4}$ quand on compare le système de RAP initial avec celui incluant des informations morpho-syntaxiques dans les scores globaux $s(w_1^n)$, $s'(w_1^n)$ ou $s''(w_1^n)$. Ceci établit donc que l'amélioration du WER est significative à 99,99 %. En outre, les mêmes tests montrent que ces trois scores permettent une meilleure prise en compte des POS que les ML N-classes au niveau du WER avec une confiance supérieure à 99,9 %. Le gain apporté par $s'(w_1^n)$ vis-à-vis de $s(w_1^n)$ est en revanche moins net puisque le *paired t-test* et le *paired Wilcoxon test* indiquent respectivement une confiance de 95 % et de 96 % sur la différence du nombre d'erreurs.

[5] Généralement fixé à 0,05 ou à 0,01.

[6] Retournant +1 si $Z_i > 0$, 0 si $Z_i = 0$ et -1 si $Z_i < 0$.

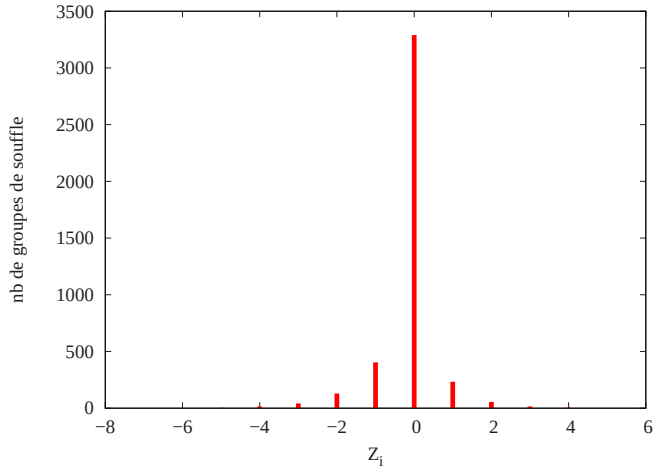

FIG. 5.3: *Calcul sur le corpus de test du nombre de groupes de souffle associés à une valeur donnée de la variable aléatoire Z_i sur le corpus de test. Le graphique présenté ici correspond à l'étude de la différence entre un système sans POS, et un utilisant ces informations à l'aide du score $s'(w_1^n)$ et du jeu d'étiquettes $ext2$*

5.2.1.4 Tests réalisés avec un meilleur MA

Nous avons également cherché à savoir si l'amélioration du WER est toujours observée avec un meilleur système de RAP. Pour ce faire, nous avons employé au niveau de la dernière passe du système IRENE un MA amélioré, utilisant des points d'ancrages [Moraru et Gravier, 2006]. Ce MA s'appuie sur une détection manuelle de la macroclasse phonétique de chaque phone à reconnaître, de façon à guider le décodage du signal. La connaissance de ces informations macrophonétiques conduit à un gain important du WER puisque celui-ci diminue de 21,6 % à 12,9 % sur le corpus de développement et de 19,9 % à 13,0 % sur les donnés de test.

La figure 5.4 présente l'évolution du WER sur le corpus de développement qui nous sert à fixer les paramètres. Elle montre là aussi une réduction du nombre d'erreurs en recourant aux POS. Nous constatons sur le corpus de test une baisse de 13,0 % à 12,1 % en utilisant $s'(w_1^n)$, ce qui correspond à une baisse absolue identique (-0,9 %) à celle obtenue avec le MA initial. Nous en concluons que même avec un système amélioré le gain apporté par les informations morpho-syntaxiques reste significatif. Il est toutefois probable qu'en modifiant le ML plutôt que le MA, la baisse n'aurait pas été aussi importante dans la mesure où les POS complémentent le ML.

5.2.1.5 Comportement sur de l'oral spontané

Bien que les informations morpho-syntaxiques aient montré au cours de nos expériences une certaine robustesse aux phénomènes de la langue parlée, il est possible qu'elles soient moins pertinentes sur les portions de la transcription qui contiennent de nombreuses disfluences. Nous souhaitons ici montrer que l'apport des POS pour la

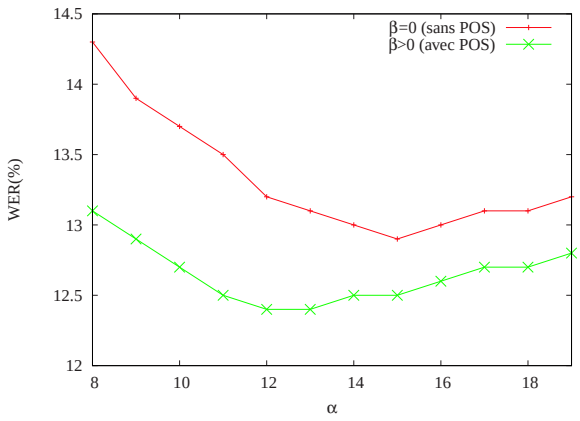

FIG. 5.4: *Évolution du WER en fonction du paramètre α avec le jeu d'étiquettes $ext2$ et sur le corpus de développement, en utilisant un meilleur MA*

	extrait de 12 640 mots	extrait de 3 650 mots
sans POS	32,7	46,3
$s(w_1^n)$	31,4	44,5
$s'(w_1^n)$	31,0	44,3

TAB. 5.5: *WER obtenus sur des extraits du corpus de test correspondant à de l'oral spontané*

transcription reste réel quand les conditions d'analyse des hypothèses de mots produites par le système de RAP se dégradent. Nous avons à cette fin isolé les extraits du corpus de test correspondant à des interviews et qui correspondent donc davantage à de l'oral spontané. Le WER initial produit par notre système de RAP sur ce passage de 12 640 mots est nettement plus élevé que les 19,9 % constatés pour la totalité des données de test (Tab. 5.5). Que l'on utilise ou non les probabilités intra-classes, les POS permettent une baisse du WER qui est en relatif identique à celle observée précédemment, voire même supérieure. Pour l'ensemble du corpus de test, nous obtenions en effet une diminution de 4,0 % ou de 4,5 % selon que l'on recourait ou non aux probabilités intra-classes ; sur ce nombre plus réduit de segments, on constate une réduction relative de 4,0 et de 5,2 %.

En examinant de plus près l'extrait que nous avons sélectionné, nous avons constaté que les portions correspondant à des interviews d'hommes politiques contenaient moins de marques de disfluences. En écartant ces groupes de souffle — ce qui restreint à un extrait de 3 650 mots — nous constatons que les POS conduisent là encore à une baisse relative du WER de 3.9 % ou de 4,3 % selon que l'on intègre ou non les probabilités intra-classes dans le calcul des scores. Ces résultats établissent par conséquent que l'utilisation des POS provoque une amélioration de la qualité de la transcription, y compris pour les zones correspondant à de l'oral spontané.

Après avoir montré l'intérêt des informations morpho-syntaxiques avec un critère MAP pour la transcription, nous proposons dans les deux sections suivantes deux nouvelles méthodes d'intégration en modifiant le critère de réordonnancement.

5.2.2 Décodage par minimisation du nombre d'erreurs sur les mots

Notre objectif étant d'améliorer la qualité de la transcription, nous proposons d'utiliser un critère plus explicite de minimisation du WER que celui consistant à maximiser la probabilité des séquences de mots proposées. Nos travaux s'inspirent de l'étude menée dans [Stolcke *et al.*, 1997]. Nous souhaitons toujours réordonner la liste des 100 meilleures hypothèses, en recourant cette fois-ci à la variable $\mathrm{WE}(W|A)$ représentant le nombre d'erreurs de transcription dans l'hypothèse proposée W à partir des observations acoustiques A. Le principe consiste à aligner entre elles les $\mathcal{N}$ hypothèses produites pour un groupe de souffle donné, selon la même méthode que celle retenue pour aligner la référence et la transcription automatique dans le calcul du WER (*cf.* section 1.3.2). La référence étant bien entendu inconnue de la part du système de RAP, le nombre d'erreurs de W_i est estimé en considérant tour à tour chaque hypothèse W_j comme la référence, ce qui conduit au calcul suivant de l'espérance :

$$E[\mathrm{WE}(W_i|A)] = \sum_{j \neq i} P(W_j|A)\mathrm{WE}(W_i|W_j) \tag{5.13}$$

où $\mathrm{WE}(W_i|W_j)$ est le nombre d'erreurs de W_i en considérant l'hypothèse W_j comme référence et $P(W_j|A)$ est la probabilité *a posteriori* de W_j. Le critère de décodage revient alors à chercher $\hat{W}$ vérifiant :

$$\hat{W} = \arg \min_{W \in W_{1,\mathcal{N}}} E[\mathrm{WE}(W|A)] \ . \tag{5.14}$$

La prise en compte des POS intervient au niveau du calcul des probabilités *a posteriori*. $P(W_j|A)$ est en effet obtenu par :

$$P(W_j|A) = \frac{P(W_j)P(A|W_j)}{P(A)} \tag{5.15}$$

où l'on reconnaît au niveau du numérateur les scores du ML et du MA, et où le dénominateur s'exprime sous la forme :

$$P(A) = \sum_k P(W_k)P(A|W_k) \ . \tag{5.16}$$

La probabilité *a posteriori* peut alors être approximée par :

$$P(W_j|A) \approx \frac{e^{score_{\mathrm{glob}}(W_j)}}{\sum_k e^{score_{\mathrm{glob}}(W_k)}} \tag{5.17}$$

		WER			SER		
		MAP	min. WE	cons.	MAP	min. WE	cons.
sans ancres	sans POS	19,9	19,8	19,8	61,8	62,2	62,4
	$s(w_1^n)$	19,1	19,0	19,1	59,9	60,1	60,7
	$s'(w_1^n)$	19,0	18,9	18,9	59,4	59,6	59,7
avec ancres	sans POS	13,0	12,9	12,8	51,3	51.8	52,7
	$s(w_1^n)$	12,3	12,2	12,1	48.6	49,2	49,8
	$s'(w_1^n)$	12,1	12,1	12,0	47,9	49,4	48,7

TAB. 5.6: *Résultats obtenus sur le corpus de test avec un système de RAP « classique », puis avec un système de RAP avec ancres*

où $score_{\text{glob}}(W_k)$ peut prendre en compte la connaissance des POS à l'aide de $s(W_k)$, $s'(W_k)$ ou $s''(W_k)$. En pratique, les scores globaux sont pondérés par un facteur z de manière à éviter des probabilités *a posteriori* trop grandes :

$$P(W_j|A) \approx \frac{e^{score_{\text{glob}}(W_j)/z}}{\sum_k e^{score_{\text{glob}}(W_k)/z}} \quad . \tag{5.18}$$

Nous observons qu'en ignorant les informations morpho-syntaxiques, ce critère de minimisation fait baisser le WER de 0,1 point par rapport à la transcription produite par le système de RAP initial. Nous constatons également qu'avec ce type de décodage, les informations morpho-syntaxiques permettent une même baisse absolue qu'avec le critère MAP (Tab. 5.6, 2 premières colonnes) ; les tests statistiques montrent que cette baisse est significative à 99,95 %.

5.2.3 Décodage par consensus

Ce troisième critère repose sur la construction d'un réseau de confusion (Fig. 1.7 page 19) à partir des 100 meilleures hypothèses produites par le système de RAP. Cette présentation de données permet d'aligner les mots de chaque hypothèse W_i en les repérant dans le groupe de souffle par une position s[7] ; en sélectionnant le mot qui apparaît le plus probable à chaque emplacement s, cette méthode permet de corriger des erreurs et de produire des sorties qui ne figuraient pas parmi les successions de mots proposées le système de RAP.

La recherche de la meilleure hypothèse est ici réalisée en examinant la variable $\text{WE}(w, s|A)$ qui représente le nombre d'erreurs (0 ou 1) faites en choisissant le mot w à la position s, pour les observations acoustiques A. La meilleure hypothèse $\hat{W}$ peut alors être déterminée en minimisant le nombre d'erreurs pour chaque position s [Mangu *et al.*, 2000], ce qui s'exprime de la manière suivante :

$$\hat{W} = \bigoplus_{s \in \mathcal{S}} \arg\min_{w \in \mathcal{W}_s} E[\text{WE}(w, s|A)] \tag{5.19}$$

[7] Pour *slot* en anglais.

où $\bigoplus$ est l'opérateur n-aire de concaténation de mots, $\mathcal{S}$ est l'ensemble des positions du réseau de confusion construit pour le groupe de souffle étudié et $\mathcal{W}_s$ l'ensemble des mots présents dans le réseau à la position s.

L'estimation de $E[\mathrm{WE}(w,s|A)]$ est réalisée à partir des probabilités *a posteriori* $P(w,s|A)$ que le mot w soit à la position s :

$$E[\mathrm{WE}(w,s|A)] = 1 - P(w,s|A) \ . \tag{5.20}$$

Le réseau de confusion étant établi à partir d'une liste des meilleures hypothèses, les probabilités *a posteriori* sont quant à elles obtenues en faisant la somme :

$$P(w,s|A) = \sum_{W \in \mathcal{H}_{w,s}} P(W|A) \tag{5.21}$$

où $\mathcal{H}_{w,s}$ est l'ensemble des hypothèses qui, une fois alignées avec le réseau de confusion (Fig. 5.5), contiennent le mot w à la position s.

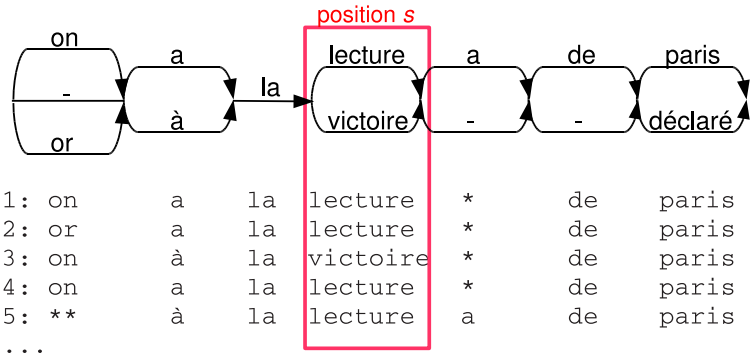

FIG. 5.5: *Alignement du réseau de confusion avec la liste des meilleures hypothèses*

Les probabilités $P(W|A)$ de chaque hypothèse W sont déterminées selon le même principe (*cf.* equation (5.18)) que le critère de minimisation explicite du WER. Nous obtenons sur le corpus de test les résultats présentés dans le tableau 5.6 (troisième colonne). L'utilisation des informations morpho-syntaxiques conduit à la même baisse absolue du WER que précédemment. En outre, les tests statistiques montrent toujours que cette amélioration est significative à 99,95 %. En revanche, que l'on utilise ou non les POS, le décodage par consensus n'apporte qu'un gain mineur par rapport aux deux autres critères. Ceci est expliqué par la taille limitée des listes d'hypothèses ($\mathcal{N}$=100) ; un gain plus important est observé quand on utilise des listes plus informatives de 1 000 hypothèses (Tab. 5.8 page 106, première ligne).

5.2.4 Une transcription facilitant l'application de méthodes du TAL

Quelle que soit la méthode de décodage utilisée, nos expériences montrent une baisse significative du WER d'environ 0,8 % en prenant en compte les informations morpho-

syntaxiques, ce qui nous permet déjà de valider notre approche. Des indices supplémentaires suggèrent que les transcriptions ainsi produites facilitent l'exploitation des sorties du système de RAP par des méthodes de TAL.

5.2.4.1 Une transcription rendue plus grammaticale

L'évolution du SER (pour *sentence error rate*) en recourant aux informations morpho-syntaxiques (Tab. 5.6, 3 dernières colonnes) montre que la diminution du WER se traduit systématiquement par une réduction du SER. En outre, en étudiant manuellement certaines erreurs corrigées par l'emploi des POS, nous avons constaté que, d'une manière générale, les groupes de souffle modifiés tendent à être plus « grammaticaux ». Cela se traduit par une baisse du nombre d'erreurs de transcription, mais aussi parfois par une apparition de mots mal reconnus. Ainsi, dans les deux premiers exemples fournis (Fig. 5.6), le recours aux informations morpho-syntaxiques permet de reconnaître correctement un passé-composé et de corriger une faute d'accord. Le dernier exemple présenté montre un cas où les POS favorisent une hypothèse syntaxiquement correcte, mais qui diffère davantage de la référence par rapport à la séquence de mots proposée initialement.

5.2.4.2 Calculs du taux d'erreur sur les lemmes

Le WER est la métrique la plus répandue pour évaluer la qualité d'une transcription. Elle présente néanmoins l'inconvénient d'accorder autant d'importance à une erreur commise sur une préposition que sur un nom qui apporte une indication importante sur le contenu du document étudié. Dans certaines applications du TAL ou de la recherche d'information, les textes sont considérés comme des « sacs de mots », en ignorant l'ordre dans lequel apparaissent les mots et en supprimant de plus les termes les plus courants, notamment ceux appartenant aux classes grammaticales ; c'est par exemple le cas des techniques de segmentation des textes suivant les sujets qu'ils traitent, application à laquelle nous nous intéressons dans le chapitre 7. Parfois, les mots sont également ramenés à des formes canoniques, telle que des lemmes, de manière à assimiler comme identiques des mots ne se distinguant que par la flexion, voire la dérivation. Nous proposons une nouvelle métrique en RAP, qui permet de mieux rendre compte que le WER de la qualité de la transcription pour ce genre d'applications, en comptabilisant le taux d'erreur sur les lemmes.

Cette mesure, notée LER pour *lemma error rate* est définie de manière similaire au WER en calculant le nombre d'insertions, de suppressions et de substitutions, mais en lemmatisant cette fois au préalable la transcription de référence et celle produite par le système de RAP. Nous annotons à cette fin ces deux transcriptions à l'aide de notre étiqueteur morpho-syntaxique, convertissons notre jeu d'étiquettes vers celui de TreeTagger, puis utilisons le lemmatiseur Flemm [Namer, 2000]. Dans la mesure où ces outils sont automatiques, des erreurs sont introduites, ce qui biaise légèrement les calculs du LER. Notre étiqueteur a néanmoins montré de bonnes performances pour annoter des émissions radiophoniques (*cf.* section 4.4.3), d'autant plus que de nombreuses er-

Erreurs corrigées	
REF	: une date qui A DONNÉ le vertige à une partie de la france
HYP(sans POS)	: une date qui À DONNER le vertige à une partie de la france
HYP(avec POS)	: une date qui a donné le vertige à une partie de la france
REF	: mais il souligne que les inspecteurs VOUDRAIENT travailler en toute indépendance
HYP(sans POS)	: mais il souligne que les inspecteurs VOUDRAIT travailler en toute indépendance
HYP(avec POS)	: mais il souligne que les inspecteurs voudraient travailler en toute indépendance
Erreurs introduites	
REF	: addition des électorats chirac balladur ELLE se CONTENTE de huit millions d' électeurs en deux mille deux
HYP(sans POS)	: addition des électorats chirac balladur **** se CONTENTENT de huit millions d' électeurs en deux mille deux
HYP(avec POS)	: addition des électorats chirac balladur PARCE QU' ON COMPTE DIX millions d' électeurs en deux mille deux

FIG. 5.6: *Exemples de groupes de souffle modifiés par la connaissance des POS*

reurs concernent des mots grammaticaux comme « *que* » simples à lemmatiser. FLEMM est de surcroît un lemmatiseur jugé performant sur le français.

L'intérêt principal du LER est qu'il ignore les erreurs d'accord ou de conjugaison qui sont peu pénalisantes pour des méthodes opérant sur des lemmes plutôt que sur des mots. En outre, les transcriptions étant étiquetées, il est possible de se limiter à certaines classes grammaticales. Seuls sont alors conservés dans les transcriptions de référence et les hypothèses produites par le système de RAP, les lemmes correspondant aux mots lexicaux que nous avons restreints aux noms communs, aux noms propres, aux adjectifs et aux verbes. Les suites de lemmes ainsi produites sont alors comparées afin de calculer une mesure que nous notons $\mathrm{LER}_{\mathrm{lex}}$. Nous définissons également $\mathrm{LER}'_{\mathrm{lex}}$ en ignorant les erreurs concernant les verbes auxiliaires les plus courants[8], qui jouent un rôle mineur dans la sémantique d'un texte.

[8]Verbes employés pour apporter des précisions sémantiques ou syntaxiques sur un autre verbe dit principal. Les verbes que nous avons considérés comme auxiliaires, quel que soit leur contexte d'utilisation, sont « *avoir* », « *être* » ainsi que les verbes modaux « *falloir* », « *devoir* », « *pouvoir* » et « *vouloir* ».

	WER	LER	LER_{lex}	LER'_{lex}
sans POS	19,9	17,2	18,5	17,5
$s(w_1^n)$	19,1	16,9	18,3	17,5
$s'(w_1^n)$	19,0	16,6	17,7	16,7

TAB. 5.7: *WER et LER (%) calculés sur le corpus de test avec un décodage par critère MAP*

La diminution relative du LER par rapport au WER mesurée sur le corpus de test (Tab. 5.7, deux premières colonnes) montrent que 13,6 % des erreurs produites par le système de RAP sans POS sont dus à des confusions sur la flexion des lemmes. Ce chiffre est à peine inférieur lorsque l'on prend en compte des informations morpho-syntaxiques (11,5 % avec $s(w_1^n)$, 12,6 % avec $s'(w_1^n)$), ce qui suggère que les POS corrigent proportionnellement un peu plus ce type d'erreurs mais qu'ils en rectifient également beaucoup d'autres. La comparaison des valeurs de LER_{lex} et LER'_{lex} suivant que l'on recoure ou non aux POS établit l'intérêt de $s'(w_1^n)$ pour améliorer la reconnaissance des lemmes. En revanche, aucun gain n'est observé pour $s(w_1^n)$ qui a tendance à privilégier les séquences de POS probables sans tenir compte des mots reconnus.

5.2.5 Application à d'autres langues que le français

Le système IRENE est conçu pour opérer sur des émissions francophones. Or, le français, en présentant notamment de nombreux homophones qui ne se distinguent que par la flexion (*cf.* section 4.1), a des caractéristiques qui laissent penser à un intérêt moindre des informations morpho-syntaxiques pour d'autres langues. Nous avons pu toutefois tester notre approche sur de l'anglais en l'appliquant en post-traitement d'un système de reconnaissance de caractères manuscrits en-ligne[9]. Ce système, développé par l'équipe IMADOC au sein de l'IRISA, calcule un score graphique et lexical en fonction des tracés des courbes constituant les caractères et de la succession des caractères formant des mots. Il recourt également à un ML bigrammes de mots construit sur un corpus de 900 000 mots. Nous l'avons étendu par un score morpho-syntaxique en adoptant une approche identique à celle que nous avons suivie en RAP. Nous avons à cette fin construit un étiqueteur basé sur les MMC à partir du corpus de Brown pour lequel un étiquetage manuel est fourni, puis calculé un score POS $P(t_1^m)$ en utilisant un lissage de Kneser-Ney et un ordre de 7. Le jeu d'étiquettes que nous avons choisi est inspiré par celui adopté dans le corpus PENN TREEBANK [Marcus *et al.*, 1993] ; nous l'avons élargi en associant des étiquettes spécifiques aux 100 mots grammaticaux les plus fréquents, ce qui conduit à un ensemble de 151 étiquettes. Le score combiné est similaire au score

[9]La qualification d'un système d'écriture de hors-ligne ou en-ligne se situe sur la manière d'acquérir les donnés. Dans des systèmes en-ligne, un stylo spécial enregistre en continu une représentation spatio-temporelle du tracé, ce qui permet de localiser clairement le signal et d'avoir des informations sur la dynamique du tracé. Dans le cas des systèmes hors-ligne, les caractéristiques du signal nécessitent une analyse du document étudié sous la forme d'une image.

$s'(w_1^n)$ calculé pour le système de RAP :

$$score = score_{\text{graphique et lexical}} + \alpha\, score_{\text{ML}} + \beta\, score_{\text{POS}} + \gamma \times \text{nb de mots} \ . \quad (5.22)$$

Ce score global est calculé pour les 100 meilleures hypothèses produites par le système de reconnaissance d'écriture manuscrite. Les coefficients α, β et γ sont optimisés en suivant un décodage MAP sur un corpus de développement de 7600 mots. La mesure du WER sur un corpus de test de 6 000 mots montre une baisse de 14,7 % à 14,3 % en prenant en compte les POS. De manière peu surprenante, l'anglais étant une langue morphologiquement moins riche que le français, nous obtenons une diminution qui est inférieure. Ce phénomène de baisse existe toutefois, alors que des expériences menées avec des ML N-classes de POS n'avaient pas permis de le constater [Quiniou *et al.*, 2005]. En outre, les paramètres α et β valant respectivement 0,3 et 0,2 sont proches, ce qui suggère que les POS sont presqu'aussi informatives que le ML de mots.

Jusqu'à présent, nous avons montré que les POS sont intéressantes pour améliorer la qualité de la transcription à partir de listes d'hypothèses. La section suivante présente un autre mode d'utilisation des POS en s'intéressant plus spécifiquement à la distinction des homophones suivant leur genre et leur nombre.

5.3 Extension de la meilleure hypothèse par ses homophones

Nous avons déjà souligné l'intérêt des POS pour corriger des confusions entre des mots homophones qui ne se distinguent que par leur flexion en genre et en nombre. Une solution peut donc consister à se restreindre à la meilleure hypothèse proposée par un système de RAP puis à corriger les erreurs dues à des confusions entre homophones pour des problèmes d'accord. Une technique similaire a d'ailleurs été employée dans une étude précédente en recourant à un graphe d'homophones pour prendre en compte la connaissance des POS en post-traitement d'un système de RAP [Gauvain *et al.*, 2005]. Nous présentons dans cette section les principes de notre méthode, avant d'exposer les résultats obtenus.

5.3.1 Principes

Afin de corriger des erreurs d'accord dans la transcription proposée par le système de RAP, nous étendons la meilleure hypothèse trouvée pour chaque groupe de souffle par une liste de ses homophones (Fig 5.7). Le nombre d'homophones pour un mot court peut être considérable puisque nous obtenons par exemple pour « *est* » 26 homophones différents selon le dictionnaire de prononciation que nous utilisons. En outre, nous nous intéressons plus particulièrement aux classes de mots susceptibles d'engendrer des erreurs d'accord. Nous avons par conséquent restreint l'extension des homophones aux adjectifs, aux noms communs, aux verbes et aux pronoms personnels. De plus, la liste des homophones des mots w_i apparaissant dans le groupe de souffle a été limitée aux

mots possédant le même lemme, *i.e.*, à ceux qui ne se distinguent que par la flexion de w_i. La suite de mots « *un texte qui finis* » se trouve à titre d'exemple étendu par « *un textes qui finis* », « *un texte qui finit* »et « *un textes qui finit* », « *un texte qui fini* »... L'ensemble des hypothèses homophones une fois obtenue, chacune est étiquetée puis réordonnée en utilisant $s'(w_1^n)$ pour introduire les POS.

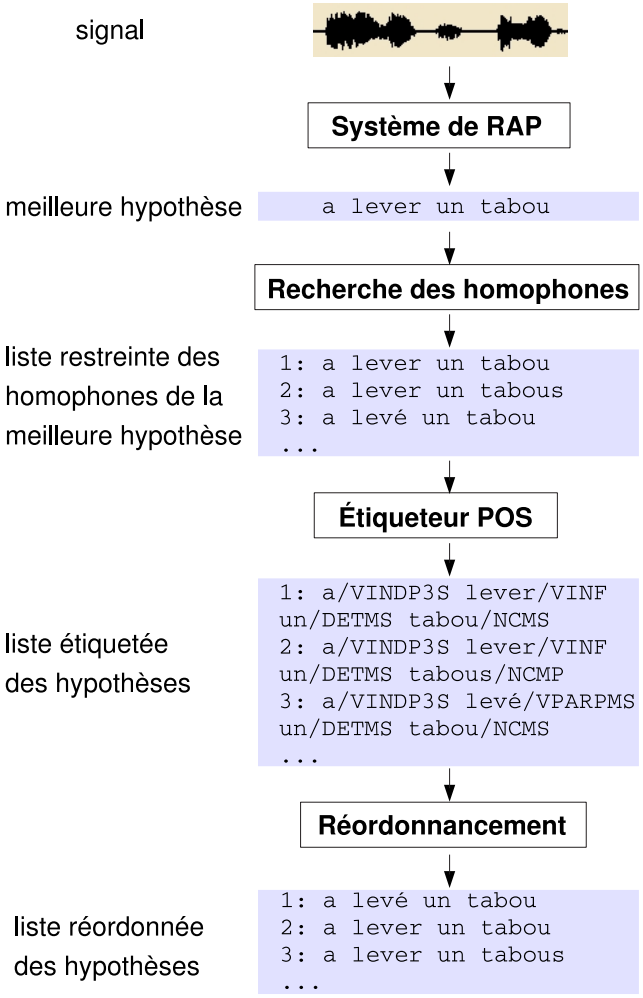

FIG. 5.7: *Principe du réordonnancement des hypothèses homophones*

5.3.2 Résultats

Nous adoptons deux stratégies pour tester l'extension des homophones. La première consiste à ne construire la liste des hypothèses en ne modifiant qu'un seul mot à la fois, en supposant que la grande majorité des groupes de souffle ne contiennent au plus qu'une erreur. En reprenant l'exemple « *un texte qui finis* » donné précédemment, ceci

permet ainsi d'éviter de considérer les séquences « *un textes qui finis* », « *un textes qui fini* »... . La seconde méthode étend simultanément plusieurs mots du groupe de souffle par un de ses homophones possibles, ce qui accroît considérablement la taille de la liste produite.

En utilisant un critère MAP, nous constatons une baisse similaire du WER avec les deux stratégies. Cette diminution est confirmée par le comportement sur le corpus de test puisque nous observons alors une diminution identique du WER de 19,9 % à 19,6 % pour les deux modes d'extension. Cette amélioration, bien qu'elle demeure significative par rapport à un système sans POS, conduit à des performances moindres que celles observées avec les 100 meilleures hypothèses, ce qui montre que la première d'entre elles corrigée par homophonie n'est pas forcément la bonne réponse. La baisse identique du WER pour les deux modes d'extension par les homophones tend en outre à privilégier la première qui conduit à réordonner en moyenne 7 hypothèses par groupe de souffle sur le corpus de développement, alors que la deuxième requiert beaucoup plus de calculs en produisant en moyenne 486 extensions par segment à analyser. Ce résultat indique également qu'il se produit généralement une erreur de confusion au plus sur le genre et le nombre par groupe de souffle.

Les deux sections précédentes traitant du réordonnancement d'hypothèses ont décrit une amélioration de la qualité de la transcription avec les POS. Nous souhaitons montrer que ce type d'information aide également à la détection des erreurs de transcription, ce qui peut permettre d'écarter des mots mal reconnus dans le cadre d'utilisation des transcriptions par d'autres applications, ou d'éventuellement appliquer des techniques de réparation locale plus coûteuses.

5.4 Détection des erreurs de transcription

Même si les objectifs sont proches, la détection des erreurs dans une transcription automatique est plus délicate à effectuer que celle réalisée par les vérificateurs orthographiques et grammaticaux intégrés dans les logiciels de traitement de texte. Une première difficulté est que les textes à analyser peuvent contenir de nombreuses erreurs successives dans un même groupe de souffle, ce qui perturbe des analyses basées sur le contexte. Heureusement, une analyse de la transcription automatique de 10 heures d'émission d'information en français a établi qu'une grande majorité des erreurs étaient isolées et que la taille des zones d'erreurs consécutives suivait grossièrement une loi de Zipf [Adda-Decker, 2006]. Une autre complication tient à la nature des documents à traiter qui présentent notamment des disfluences, sujettes à entraîner des ruptures dans la séquence de mots à analyser. En outre, compte tenu de la nature des erreurs de transcription qui se produisent entre des mots homophones, plutôt que des mots homographes, les hypothèses erronées peuvent avoir une écriture très éloignée du mot correct, ce qui empêche le recours à des méthodes basées sur la proximité lexicale. Il est d'ailleurs à noter que les coquilles de saisie, tel que l'emploi de « *abre* » en lieu et place de « *arbre* », sont absentes des transcriptions produites automatiquement puisque

les systèmes de RAP ne reconnaissent que les mots de leur vocabulaire. Une part importante des erreurs de reconnaissance, comme l'a montré notre analyse d'un extrait de transcription (*cf.* section 4.1), sont en revanche des fautes d'orthographe grammaticales, liées à des erreurs d'accord ou de conjugaison et pour lesquelles l'exploitation de la connaissance des POS constitue un indice de détection. Nous montrons ici que les informations morpho-syntaxiques sont pertinentes pour repérer les erreurs de transcription globalement, *i.e.*, au niveau d'un groupe de souffle entier et surtout localement, *i.e.*, au niveau d'un mot. Nous utilisons cette propriété pour calculer des mesures de confiance qui se révèlent meilleures que lorsque nous ignorons les POS.

5.4.1 Détection globale

Nos expériences sur le réordonnancement des hypothèses montrent que les POS tendent à privilégier celles qui possèdent moins d'erreurs d'accord (*cf.* section 5.2.4). Dans le but de mesurer l'efficacité du score morpho-syntaxique $P(t_1^m)$ pour détecter les hypothèses présentant ces erreurs, nous examinons son comportement sur les fautes d'accord que nous avons repérées dans l'extrait examiné manuellement (*cf.* section 4.1) et pour lesquelles les informations morpho-syntaxiques semblent particulièrement pertinentes. Pour chacun des groupes de souffle contenant au moins une erreur d'accord interne au groupe de souffle et qui ne concerne pas de mots absents du dictionnaire de prononciation, nous établissons le score pour trois versions : la transcription de référence (REF), la transcription automatique (HYP) et la transcription automatique où seules les erreurs d'accord sont corrigées par rapport à HYP (COR) (Fig. 5.8). Nous espérons ainsi que la succession d'étiquettes obtenues sur REF et COR sera plus probable que sur HYP.

```
REF: à  L' AMÉNAGER avant qu' elle ne soit DÉTRUITE
COR: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUITE
HYP: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUIT
```

FIG. 5.8: *Versions à évaluer pour un même groupe de souffle*

En utilisant notre jeu d'étiquettes *norm* de 93 étiquettes, nous avons constaté sur les 62 groupes de souffle analysés que le score était strictement supérieur sur COR par rapport à celui sur HYP pour 42 d'entre eux et strictement plus grand sur REF que sur HYP pour 31 d'entre eux. Sans surprise, $P(t_1^m)$ permet de favoriser les groupes de souffle sans erreurs d'accord. Il offre des informations supplémentaires à celles apportées par le ML de mots 4-grammes utilisé au sein du système de RAP puisque sur ce même ensemble de 62 groupes de souffle, le score $P(w_1^n)$ qu'il calcule est strictement plus grand sur COR (resp. sur REF) que sur HYP pour 6 (resp. 14) d'entre eux. Notons que les erreurs commises par le score morpho-syntaxique sur la discrimination de COR et de HYP sont dues à des erreurs d'étiquetage ou à la nécessité de faire une analyse syntaxique plus poussée comme dans les exemples de la figure 5.9. La dégradation importante des résultats entre COR et REF montre néanmoins qu'il faut plutôt envisager une approche locale pour repérer des erreurs, en particulier les fautes d'accord.

```
REF: pas question ****** RÉPONDENT l' union européenne et la russie qui
     REFUSENT de lever les sanctions
HYP: pas question RÉPOND DE       l' union européenne et la russie qui
     REFUSE   de lever les sanctions
REF:  les supporters de l' om APPRÉCIENT thibaud vincent
CORR: les supporters de l' om APPRÉCIENT thibaud vincent
HYP:  les supporters de l' om APPRÉCIE   thibaud vincent
```

FIG. 5.9: *Exemples d'erreurs d'accord nécessitant une analyse syntaxique*

5.4.2 Détection locale

Nous nous intéressons désormais à une détection plus fine des mots erronés. Nous utilisons à cette fin les probabilités contextuelles $P(t_i|t_{i-N+1}^{i-1})$, où les t_i sont les étiquettes produites par notre étiqueteur, en espérant que leur valeur décroît fortement quand le mot associé à t_i est erroné. Une analyse de l'évolution de $P(t_i|t_{i-N+1}^{i-1})$ pour chaque mot d'un groupe de souffle contenant une erreur d'accord en genre montre ainsi une baisse importante au niveau du mot erroné (Fig. 5.10). En comparant les résutats obtenus avec plusieurs méthodes de lissage, celles de Kneser-Ney et de Ristad en l'occurrence, on constate peu de différences avec la courbe obtenue sans effectuer de lissage dans le calcul des probabilités. De la même façon, l'ordre N semble avoir peu d'influence. Le rapprochement des courbes obtenues avec des ML 7-grammes à base de POS avec celles produites par un ML quadrigrammes de mots montre en revanche des différences de comportement puisque celles construites par des ML de POS sont globalement plus lisses, exceptées au niveau de l'erreur d'accord. Ceci n'est pas surprenant dans la mesure où le nombre des N-grammes possibles est beaucoup plus grand dans le cas des ML de mots, ce qui fait que de nombreuses séquences sont rares ou peu observées et ont des probabilités contextuelles faibles incluant des coefficients de repli. Cette constatation laisse donc espérer une plus grande capacité des ML à base de POS pour détecter certaines erreurs de transcription.

Afin de valider la capacité des POS à détecter localement les mots mal reconnus, nous recourons à un seuil relatif δ_{rel}, en espérant les probabilités contextuelles sont inférieures à cette valeur pour des mots effectivement mal reconnus. Nous considérons alors que le mot w_i étiqueté par t_j est erroné lorsque :

$$\log P(w_i|w_{i-M+1}^{i-1}) + \eta \times \log P(t_j|t_{j-N+1}^{j-1}) \leq \mu + \sigma \times \delta_{\text{rel}} \tag{5.23}$$

où μ et σ sont respectivement la moyenne et l'écart-type de la combinaison des log-probabilités conditionnelles déterminées sur un groupe de souffle donné.

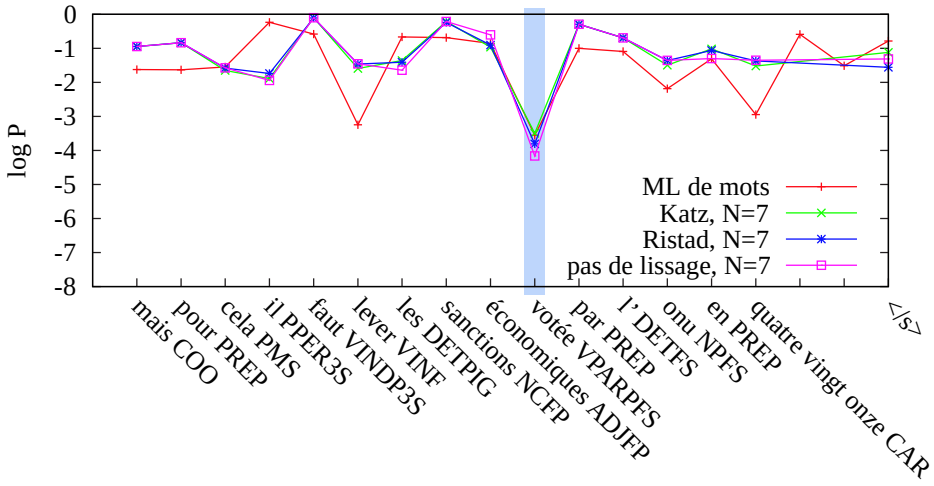

FIG. 5.10: *Courbe* $\log P(t_i|t_{i-N+1}^{i-1}) = f(w_i)$ *tracée pour une hypothèse présentant une erreur d'accord sur « votée », en utilisant le jeu d'étiquettes* $norm$

Nous mesurons la capacité à repérer les erreurs en calculant des valeurs de rappel[10] et de précision[11] sur le court extrait que nous avons analysé manuellement. Nous nous focalisons ici sur la détection de l'ensemble des erreurs de transcription et sur celle plus spécifique des erreurs d'accord corrigeables en examinant isolément chaque groupe de souffle. Le ML de mots employé est celui utilisé dans la troisième passe du système IRENE et correspond ainsi à un ordre $M = 4$, tandis que notre ML de classes est construit sur le jeu d'étiquettes *norm* avec un ordre $N = 7$ et un lissage de Katz. Le graphe rappel-précision présenté à la figure 5.11 montre les résultats obtenus en faisant varier le seuil δ_{rel}. Afin d'étudier l'influence des POS sur la détection, nous comparons les courbes pour trois valeurs différentes de η, en fixant notamment ce coefficient de pondération à 0 de manière à ne prendre en compte que le ML de mots. Nous constatons alors que l'utilisation des POS conduit à améliorer la précision pour la plupart des valeurs de rappel par rapport au ML de mots seul (η=0), notamment pour les valeurs les plus élevées du seuil. On peut en outre remarquer que l'apport des POS est un nettement plus marqué pour les erreurs d'accord. Néanmoins, les résultats sont globalement décevants puisque la précision diminue rapidement quand on augmente le rappel. L'apport des POS devient de plus négligeable lorsque le rappel devient supérieure à 50 %. Les valeurs de précision mesurées pour les fautes d'accord sont particulièrement basses, du fait que la méthode de détection retourne principalement des mots mal reconnus

[10] Proportion des erreurs correctement détectées par rapport à l'ensemble des erreurs de transcription dans le cas général ; proportion des erreurs d'accord correctement détectées par rapport à l'ensemble des fautes d'accord internes au groupe de souffle, en se limitant aux accords non reconnus.

[11] Proportion des erreurs correctement détectées parmi les réponses présentées comme des erreurs par le système dans le cas général ; proportion des erreurs d'accord correctement détectées parmi les réponses présentées comme des erreurs par le système, en se limitant aux accords non reconnus.

pour d'autres raisons. Ces observations nous conduisent à combiner davantage d'informations, en recourant notamment aux indices acoustiques, pour obtenir des résultats plus satisfaisants, et à employer plutôt la probabilité *a posteriori* que la probabilité contextuelle.

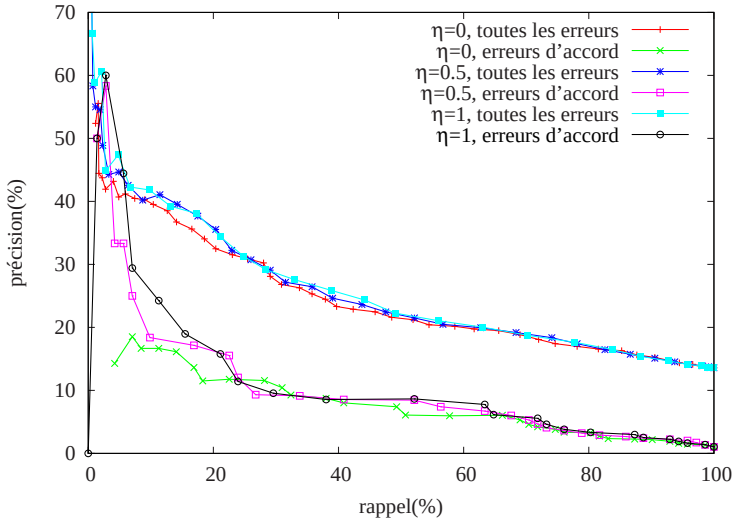

FIG. 5.11: *Courbe rappel-précision sur la détection des erreurs de transcription par un ML de mots et un ML de POS basé sur le jeu d'étiquettes norm*

5.4.3 Mesures de confiance

Nous avons montré précédemment l'intérêt des POS pour calculer des probabilités *a posteriori* d'hypothèses dans le cadre des critères de minimisation du nombre d'erreurs de transcription (*cf.* sections 5.2.2 et 5.2.3). Ces probabilités sont communément employées pour établir des mesures de confiance à partir de listes d'hypothèses ou de graphes de mots [Wessel *et al.*, 2001]. Nous recourons aux POS pour améliorer ces mesures, celles-ci pouvant fournir une indication sur la probabilité qu'un mot soit erroné ou non, et pouvant être employés par des applications qui utilisent des transcriptions, en écartant les mots qui risquent d'avoir été mal reconnus.

Nous calculons les mesures de confiance à partir des listes d'hypothèses produites avec un décodage MAP ou des réseaux de confusion établis dans le cadre du décodage par consensus (Fig. 5.12). Dans le premier cas, les mesures de confiance sont établies à partir des probabilités *a posteriori* calculées au niveau du groupe de souffle. Suivant la méthode proposée dans [Rueber, 1997], les hypothèses de chaque liste sont alignées avec la meilleure. La mesure de confiance pour un mot w_i situé à la position s de cette meilleure hypothèse est alors la somme des probabilités *a posteriori* des groupes de souffle qui contiennent également w_i à cette même position. Dans le cas de réseaux

de confusion, les mesures de confiance sont directement données par la probabilité *a posteriori* la plus élevée pour chaque position s du réseau de confusion.

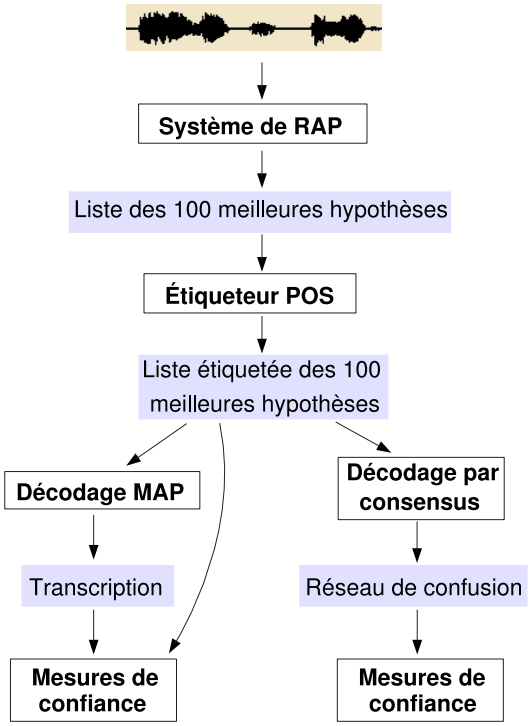

FIG. 5.12: *Principe du calcul des mesures de confiance avec un décodage MAP ou un décodage par consensus*

Afin d'évaluer la qualité des mesures de confiance obtenues, nous recourons à la mesure d'entropie croisée normalisée, notée NCE pour *normalized cross entropy*. Cette métrique, introduite par l'agence NIST, estime l'information mutuelle entre les mots correctement transcrits par le système de RAP et les mesures de confiance calculées, en la normalisant par l'entropie croisée maximale :

$$NCE = [H_{\max} + \sum_{w \text{ corrects}} \log(\hat{p}(w)) + \sum_{w \text{ incorrects}} \log(1 - \hat{p}(w))]/H_{\max} \tag{5.24}$$

avec :

$$H_{\max} = -n_c \log(p_c) - (n_{all} - n_c) \log(1 - p_c) \tag{5.25}$$

$$p_c = \frac{n_c}{n_{all}} \ . \tag{5.26}$$

n_c est le nombre de mots correctement transcrits, n_{all} le nombre total de mots de la transcription automatique et $\hat{p}$ la mesure de confiance à évaluer. Ainsi, plus la NCE est élevée, meilleure est considérée la mesure de confiance.

Les mesures de confiance sont calculées à partir de listes de 1000 hypothèses, plus informatives que celles de 100 précédemment utilisées. Ces probabilités sont calculées à

	décodage MAP		décodage par consensus	
	sans POS	avec POS	sans POS	avec POS
WER	19,7	18,7	19,4	18,6
NCE sans POS	0,307	0,265	0,198	-
NCE avec POS	0,326	0,288	-	0,211

TAB. 5.8: *WER (%) et NCE calculés sur le corpus de test pour différentes méthodes de décodage*

partir du score $s'(w_1^n)$, en utilisant le jeu d'étiquettes $ext2$ employé dans les expériences sur le réordonnancement. Dans le cas du décodage MAP, les facteurs de pondération α, β, γ et z intervenant dans le calcul des probabilités *a posteriori* sont optimisés spécifiquement sur le corpus développement pour maximiser la NCE. Dans le cas du décodage par consensus, les réseaux de confusion utilisés ont déjà été optimisés pour minimiser le WER.

Le tableau 5.8 expose au niveau de la première ligne les WER mesurés suivant différentes configurations, *i.e.* selon un décodage MAP ou par consensus, avec ou sans POS. Les deux suivantes affichent les valeurs de la NCE calculées sur le corpus de test. Ces résultats établissent clairement l'intérêt des POS pour améliorer le calcul des mesures de confiance. En effet, avec un décodage MAP, les POS apportent un net gain en les employant dans le calcul des probabilités *a posteriori*, qu'elles aient été auparavant utilisées durant le décodage (0,326 par rapport à 0,307) ou pas (0,288 par rapport à 0,265). Dans le cas du décodage par consensus, l'apport des POS est plus délicat à montrer dans la mesure où l'on ne peut comparer facilement des mesures de confiance obtenues pour des hypothèses de transcription avec des WER différents. Toutefois, une hausse de la NCE peut là encore être observée en recourant aux POS. Cette augmentation est moins importante que pour le premier décodage car les paramètres ne sont pas optimisés spécifiquement pour les mesures de confiance.

L'examen des courbes de détection d'erreur obtenues avec un décodage MAP avec POS sur le corpus de développement montre en outre que l'amélioration constatée en utilisant des informations morpho-sytaxiques concerne surtout les hautes valeurs des mesures de confiance, *i.e.*, quand seules les mots très probablement erronés sont écartés (Fig. 5.13). En utilisant des POS et en excluant tous les mots dont la mesure de confiance est inférieure à un certain seuil, il est par exemple possible d'éliminer 5 % des mots erronés de la transcription, en ôtant seulement 0,1 % des mots corrects. Si l'on n'utilise plus cette connaissance, on voit que l'on écarte près de 0,4 % de mots corrects si l'on souhaite éliminer 5 % de mots erronés.

5.5 Bilan

Ce chapitre établit clairement l'intérêt des informations morpho-syntaxiques pour réduire le nombre de mots mal reconnus. Nous avons défini un nouveau score incluant la connaissance des POS en post-traitement d'un système de RAP et permettant une

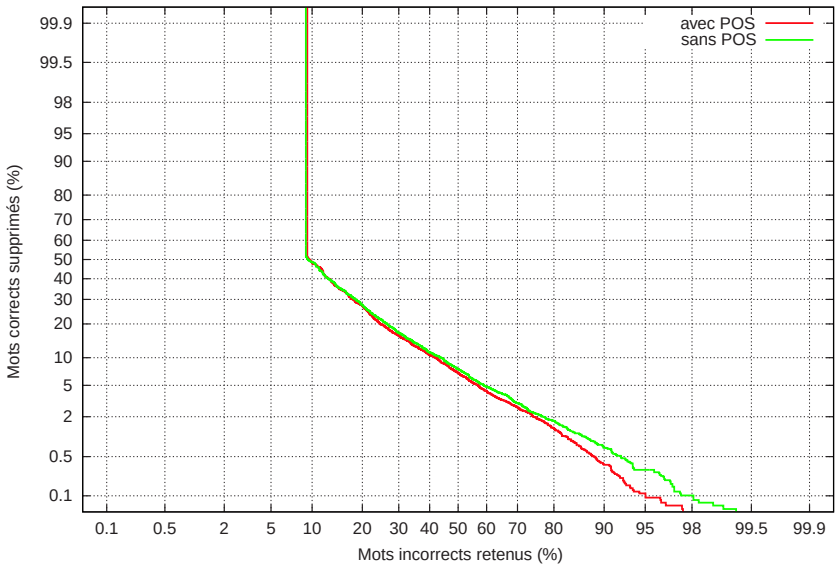

FIG. 5.13: *Courbes de détection d'erreurs obtenues sur le corpus de développement suivant un décodage MAP avec POS (WER=20,2 %, NCE sans POS=0,223, NCE avec POS=0,241). Ces courbes sont construites en faisant varier un seuil sur le score MAP*

diminution absolue du WER de 0,9 % pour transcrire des émissions d'information. Cette baisse, que nous avons prouvée statistiquement significative, a été vérifiée dans différentes conditions, que ce soit en utilisant un MA plus informatif ou sur des extraits contenant davantage de disfluences. Nous avons de surcroît recouru à trois critères différents : le critère MAP, la minimisation explicite du WER et le décodage par consensus, qui ont tous indiqué un apport des POS pour améliorer la qualité de la transcription. Une expérience complémentaire menée avec un système de reconnaissance d'écriture manuscrite laisse à penser que ces résultats sont extensibles sur d'autres langues que le français. Nous avons enfin montré la capacité des POS à détecter des erreurs de transcription, en constatant notamment que leur emploi conduit à une amélioration des mesures de confiance. Cette propriété est intéressante pour les applications qui utilisent les productions d'un système de RAP puisqu'elle permet d'écarter certains mots probablement mal reconnus.

Les réductions du SER et du LER signalent que les transcriptions, obtenues lorsque l'on tire profit, grâce à notre score original, des POS, sont plus grammaticales et comportent moins d'erreurs sur les mots lexicaux. Nous nous appuyons dans le chapitre suivant sur cette propriété pour employer des méthodes de TAL concernant la segmentation des émissions transcrites en fonction des thèmes abordés et l'extraction des mots-clés. Ces techniques sont utilisées dans le cadre de l'adaptation thématique de notre système de RAP.

Troisième partie

Adaptation thématique de la reconnaissance de la parole

Chapitre 6

Chaîne de traitements

Dans la dernière partie de ce mémoire, nous nous intéressons à la spécialisation d'un système de RAP en fonction des sujets abordés par le document étudié. Notre but est donc là bien encore de recourir à des méthodes de TAL pour améliorer la transcription. L'intérêt de ces travaux est de pallier le fait que les systèmes de RAP actuels utilisent le même ML général pour traiter chaque document qu'ils rencontrent. Bien que ce ML soit construit à partir d'un corpus de taille volumineuse, cet ensemble de données ne peut prétendre représenter l'intégralité d'une langue, d'autant plus que la fréquence d'utilisation des expressions et des mots évolue au cours du temps. Cette constatation peut être particulièrement faite dans le cas des émissions d'information qui voient souvent apparaître et disparaître des entités nommées, notamment des noms de personnes, suivant les événements marquants de l'actualité. Un autre inconvénient de ce type de ML est qu'il est appliqué uniformément sur l'intégralité d'un document, bien que celui-ci puisse aborder des thèmes très divers, avec un vocabulaire et des modes de formulation particuliers. Un reportage traitant de la guerre civile en Côte d'Ivoire utilise ainsi fréquemment des mots se rapportant à la guerre, en nommant les belligérants du conflit, tandis qu'une revue de presse sur l'actualité économique contient de nombreux chiffres et noms d'entreprises.

La modification du système de RAP est souhaitable à plusieurs niveaux pour traiter de la variabilité des documents étudiés. Afin de prendre en compte les évolutions des conditions acoustiques d'enregistrement ou encore l'apparition de nouveaux locuteurs, il est ainsi souhaitable de modifier le MA. Le vocabulaire gagnerait également à être transformé puisque des domaines différents peuvent avoir des mots spécifiques. Enfin, il est avantageux de modifier le ML pour modéliser spécifiquement les particularités du langage de chaque thème du document à analyser. Dans ce mémoire, nous nous limitons à la mise à jour du ML. Si l'extension du vocabulaire est intéressante, elle soulève en effet plusieurs difficultés telles que le choix des mots à ajouter, celui de ceux à supprimer pour ne pas avoir un lexique trop important, l'écriture de règles de conversion automatique de mots en phonèmes [Yvon *et al.*, 1998] et la modification des probabilités du ML pour les nouveaux mots [Allauzen et Gauvain, 2005]. La modification du MA sort quant à elle du cadre que nous nous sommes fixé dans nos travaux ; elle est en outre prise en

compte par de nombreux systèmes — dont le nôtre — qui possèdent déjà un mécanisme d'adaptation au locuteur et des modèles particuliers à certaines conditions acoustiques comme le téléphone.

Au cours de nos travaux, nous avons cherché à étudier si la connaissance du sujet traité dans une partie donnée d'un document permet d'améliorer la transcription. La détection de thème est en effet un domaine déjà bien étudié dans le domaine du TAL et nous avons voulu savoir si certaines de ces techniques pouvaient s'appliquer à des transcriptions automatiques. L'utilisation de ces méthodes dans ce cadre pose plusieurs difficultés du fait de la nature des documents traités et de l'imprécision sur ce que représente un thème. Les travaux que nous menons dans cette partie sont donc plus exploratoires que dans la précédente et visent avant tout à examiner ce que pourrait apporter cette notion de thème pour la RAP, en posant des points de repère pour des recherches futures.

Nous contribuons au domaine de l'adaptation thématique en proposant une chaîne de traitements qui débute par une segmentation en fonction des sujets abordés d'une première transcription produite avec un ML général, et qui s'achève par l'application de ML mis à jour pour chaque section à partir de corpus récupérés sur Internet. Peu de travaux à notre connaissance recourent à une chaîne aussi complète pour adapter le système de RAP. Une autre particularité de nos travaux est que nous nous intéressons à la segmentation thématique d'une transcription automatique, en vue de l'adaptation du ML, alors que ce type de découpage est généralement envisagé dans d'autres cadres applicatifs. Afin de constituer nos corpus d'adaptation, nous recourons en outre à des systèmes de recherche d'information (RI) et au Web, dont l'utilisation reste encore marginale pour la formation de ML. Cette manière de procéder nous permet de réaliser une adaptation non supervisée, en ne faisant aucune hypothèse sur les thèmes traités par les documents étudiés.

Dans la suite de ce chapitre, nous exposons les travaux qui présentent des similarités avec les nôtres, avant de passer en revue chacun des composants de notre chaîne de traitements.

6.1 Travaux connexes

Nous nous limitons dans cette section à la synthèse des travaux connexes qui s'intéressent à l'adaptation de ML au sens large, à la sélection de ML et de corpus pertinents pour chaque document traité, à l'emploi du Web pour modéliser le langage et enfin à ceux qui effectuent un découpage des transcriptions avant d'appliquer des ML spécialisés à chaque section identifiée. Tout ce qui touche à la segmentation thématique proprement dite est présentée en détail dans le chapitre 7.

L'adaptation des ML est étudiée depuis une dizaine d'années [Bellegarda, 2004]. Elle est principalement utilisée pour modifier un ML appris sur un corpus général, en fonction d'un petit volume de données spécifiques au type de documents que l'on souhaite transcrire. Ceci correspond typiquement au cas de figure où l'on ne dispose pas de suffisamment d'heures transcrites manuellement pour la tâche visée, à cause de l'ef-

fort important que cela demande. Des corpus constitués de documents écrits peuvent ainsi apporter une bonne couverture du langage, tandis qu'un corpus oral plus court est chargé d'apprendre les formes d'expression les plus courantes pour les documents à traiter. L'adaptation d'un système de RAP en fonction des thèmes effectivement abordés dans une transcription est beaucoup moins étudiée. Nous avons déjà décrit en section 2.2.5 les principales méthodes employées à cette fin, notamment les modèles thématiques. Le point délicat à résoudre pour ces types de modèles concerne le choix des thèmes ; même si ceux-ci sont généralement fixés automatiquement en partitionnant un corpus en groupes de documents similaires, le choix du nombre de sujets reste problématique, d'autant plus que les thèmes ne sont souvent pas caractérisés et qu'il est donc difficile de savoir à quoi ils correspondent. Un autre inconvénient est que cette méthode est dépendante du corpus d'apprentissage. De nouveaux sujets, absents de la liste initiale des thèmes, peuvent ainsi apparaître régulièrement. En outre, les documents traitent souvent plusieurs sujets, ce qui demande de savoir combiner différents modèles thématiques.

Afin d'adapter spécifiquement le ML à chaque document à traiter, plusieurs solutions sont envisageables. Si des modèles thématiques ont été préalablement construits, des méthodes de catégorisation de textes [Sebastiani, 2002], dont en particulier celles de détections de thèmes [Brun, 2003], peuvent être employées dans ce cadre. Parfois, le document à transcrire est accompagné de méta-données indiquant un résumé, le genre auquel il appartient ou la date à laquelle il a été produit. Ces informations peuvent alors être utilisées pour rechercher des textes présentant des similitudes avec les documents étudiés [Allauzen et Gauvain, 2005]. Une dernière solution consiste à recourir aux techniques de RI [Chen *et al.*, 2001 ; Bigi *et al.*, 2004] pour rechercher dans une base de données les textes qui se rapportent aux phrases déjà reconnues.

Le Web est devenu une ressource utilisée par un nombre important de tâches du domaine du TAL [Kilgarriff et Grefenstette, 2003 ; Lapata et Keller, 2005] telles que la désambiguïsation sémantique, la correction contextuelle de fautes d'orthographe ou encore la traduction automatique. Il est également employé pour construire des ML, en téléchargeant des pages Web pour construire des corpus d'apprentissage [Berger et Miller, 1998 ; Nisimura *et al.*, 2001 ; Vaufreydaz, 2002 ; Bulyko *et al.*, 2003 ; Sethy *et al.*, 2005 ; Suzuki *et al.*, 2006] ou bien en estimant la fréquence d'un N-grammes donné à partir du nombre de pages retournées par un moteur de recherche en résultat d'une requête formée de ce N-grammes [Zhu et Rosenfeld, 2001]. Le recours au Web pour établir des ML revêt plusieurs intérêts. Il s'agit tout d'abord d'une vaste source d'information — des milliards de pages étant référencées par les moteurs de recherche — qui est en outre continuellement mise à jour, ce qui permet de constituer des corpus d'adaptation même pour des événements récents. Certains documents, tels les *blogs*, ont de plus un langage plus proche des corpus oraux que beaucoup d'autres textes de la langue écrite, ce qui les rend intéressants pour construire des ML [Vaufreydaz *et al.*, 1999]. Cette grande diversité est néanmoins également problématique, puisque les formats à traiter peuvent être différents et que certains documents sont de mauvaise qualité pour modéliser le langage. Pour ces raisons, les ressources utilisées ne sont parfois

limitées qu'à certains types de pages, comme celles apparaissant sur des sites d'archives ou de quotidiens [Kemp et Waibel, 1998 ; Scheytt *et al.*, 1998 ; Allauzen, 2003].

Au final, l'utilisation d'une chaîne complète de traitements, depuis la segmentation thématique jusqu'à la spécialisation spécifique du ML en fonction de chaque partie identifiée, reste rare. Nous n'avons rencontré ainsi que trois études qui effectuent un découpage du document transcrit, puis une adaptation du ML — sans recourir toutefois comme nous le faisons au Web. Les travaux de Chen *et al.* [2001], poursuivis dans [Chen *et al.*, 2003] et [Chen *et al.*, 2004a], sont ceux qui se rapprochent le plus des nôtres. Les auteurs s'intéressent au même cadre applicatif que le nôtre, *i.e.*, à la transcription d'émissions d'actualité. Ils recourent de même à des méthodes de la RI pour former leurs corpus d'adaptation. Néanmoins, leur segmentation est basée essentiellement sur des indices acoustiques et son influence sur l'amélioration du WER n'est pas étudiée. En outre, les données à partir desquelles sont adaptés les ML sont issues d'un corpus constitué d'émissions d'information et fixé au départ.

Sako *et al.* [2007] étudient quant à eux l'adaptation du ML pour transcrire des commentaires de matchs de base-ball. Le calcul des probabilités par le ML modifié s'effectue en même temps que la segmentation en recherchant le thème associé à chaque groupe de souffle. Si t_i représente le sujet abordé par le groupe de souffle contenant le mot w_i et $T = t_1 \dots t_n$ est la succession des thèmes traités par l'ensemble du document à transcrire $W = w_1^n$, la probabilité des mots transcrits est alors calculée par :

$$P(W) = \sum_T P(W,T) \approx \max_T \prod_{i=1}^{n} P(t_i|t_{i-1}) P(w_i|w_{i-2}^{i-1}, t_i) \tag{6.1}$$

où $P(t_i|t_{i-1})$ est la probabilité de transition entre deux thèmes, égale à 1 quand w_{i-1} ne correspond pas à une fin de groupe de souffle, et $(w_i|w_{i-2}^{i-1}, t_i)$ est calculée à partir d'un modèle trigrammes, adapté pour correspondre au thème t_i. Les inconvénients principaux de cette méthode sont que l'évaluation de $P(t_i|t_{i-1})$ est très dépendante du corpus étudié et que le nombre de thèmes possibles est difficile à fixer expérimentalement.

Enfin, les travaux d'Akita *et al.* [2007], s'intéressant à la transcription d'échanges prononcés dans le cadre de réunions, utilisent une méthode en ligne pour segmenter le document à transcrire. La section courante est étendue jusqu'à ce que qu'elle devienne trop différente du groupe de souffle courant à reconnaître et un nouveau segment débute alors. Les documents les plus similaires parmi un corpus sont employés pour adapter le modèle à chaque section. La méthode proposée contient deux seuils délicats à régler : l'un pour décider des lieux de rupture thématique, et l'autre pour choisir si un document est similaire à la section étudiée et doit appartenir ou non au corpus d'adaptation.

Après avoir passé en revue les études comparables à nos travaux, nous présentons dans la section suivante l'approche que nous avons choisie et les arguments de ces choix.

6.2 Notre approche

Notre système d'adaptation est constitué de plusieurs étapes successives (Fig. 6.1). Tout au long des traitements que nous effectuons, nous limitons le nombre de connais-

sances supposées sur les documents à transcrire. Nous avons ainsi voulu ne pas nous restreindre à une liste prédéfinie de thèmes et avons souhaité constituer des corpus d'adaptation à partir d'un ensemble de textes qui ne reste fixe pour tous les documents traités.

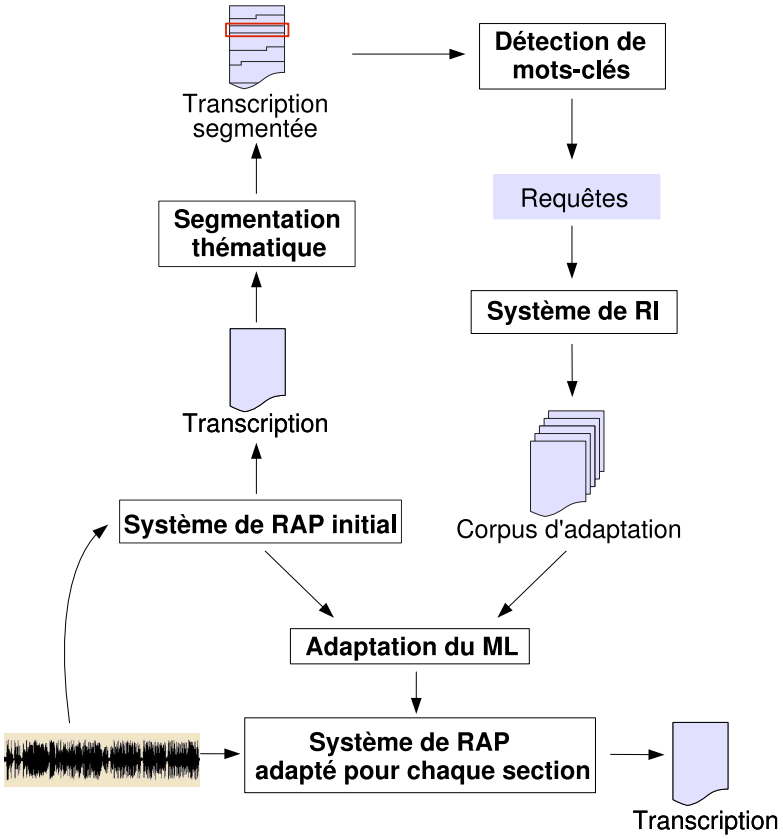

FIG. 6.1: *Principe de l'adaptation du système de RAP pour chaque section thématique*

La première étape de l'adaptation de notre système de RAP consiste à identifier les ruptures thématiques dans une transcription automatique. Nous nous appuyons principalement sur l'étude de la distribution des mots porteurs de sens au sein du document à transcrire. Nous proposons une nouvelle méthode qui permet de combiner ce type d'informations avec d'autres basées sur le repérage de marqueurs linguistiques et la recherche d'indices acoustiques signalant une rupture thématique. La principale difficulté rencontrée concerne le choix du niveau de segmentation pour adapter convenablement le ML, les techniques de détection de rupture thématique n'étant que très rarement employées dans notre cadre applicatif. La segmentation thématique fait l'objet du chapitre 7.

Une deuxième étape vise à récupérer un corpus d'adaptation pour chaque segment précédemment identifié. Pour ce faire, nous identifions les mots jugés les plus repré-

sentatifs de chaque section, en recourant à des méthodes usuelles de RI. Ces mots-clés sont ensuite utilisés pour formuler des requêtes à un moteur de recherche sur Internet. Les pages Web proposées dans les premières pages de réponse obtenues pour chaque requête permettent alors de constituer des corpus d'adaptation. Cette méthode, peu employée en RAP pour transcrire des documents portant sur des thèmes très variés, nous permet d'avoir un ensemble de données dont on peut espérer qu'il traite différents aspects du sujet abordé. Elle suppose toutefois de développer de nouvelles procédures pour s'assurer que les documents constituant le corpus d'adaptation soient en nombre suffisant et correspondent bien aux sections transcrites, et ce quel que soit le sujet dont il est question.

Dans une dernière étape, nous modifions les probabilités de notre modèle N-grammes en fonction du corpus d'adaptation trouvé pour chaque section thématique. Nous n'avons envisagé pour l'instant qu'une méthode simple d'adaptation, à savoir l'interpolation linéaire du ML général avec un ML spécifique. L'objectif de notre travail est en effet de poser les bases d'un système d'adaptation du ML et cette étape finale doit plutôt être vue comme un moyen d'évaluer l'intérêt de notre approche pour améliorer la qualité de la transcription. La constitution de corpus d'adaptation ainsi que la modification du ML font l'objet du chapitre 8.

Nous nous intéressons donc maintenant plus précisément à la première opération de notre chaîne de traitements, *i.e.*, la segmentation d'une transcription suivant les thèmes qu'elle aborde.

Chapitre 7

Segmentation thématique

La première étape de notre chaîne de traitements consiste, comme nous venons de le voir, à segmenter selon les thèmes abordés la transcription produite par notre système de RAP. Cette opération vise à obtenir des sections homogènes de manière à pouvoir adapter plus finement le système de transcription.

La tâche de segmentation que nous visons pose plusieurs difficultés. La première d'entre elles, commune à tout type de texte, concerne la polysémie, *i.e.*, la possibilité pour un mot d'être employé avec plusieurs sens distincts. Cette propriété du langage complique singulièrement le repérage des changements de sujets puisque plusieurs occurrences d'un même mot peuvent être associées à des thèmes différents. Un autre obstacle, lié au genre des documents étudiés, est la présence de mots mal reconnus qui représentent autant de bruit, propre à mettre en erreur les décisions de positionnement des frontières thématiques. Enfin, la notion de thème pouvant mener à différentes interprétations, le type de découpage d'un document à obtenir pour améliorer effectivement la qualité de la transcription n'est pas clairement établi.

Nous nous intéressons au domaine de la segmentation thématique dans un cadre applicatif qui demeure encore largement inexploré. Peu de travaux cherchent en effet à partitionner des transcriptions automatiques et le découpage d'un document en vue d'une adaptation à chaque sujet reste rare pour transcrire de la parole. La méthode de segmentation sur laquelle nous basons nos travaux est celle proposée dans [Utiyama et Isahara, 2001], qui a déjà montré de bonnes performances sur des documents de langue écrite et dont le principe est applicable à n'importe quel type de texte. Nous étudions l'apport de connaissances supplémentaires à cette méthode reposant uniquement sur la répétition des mots, afin d'améliorer la qualité de la segmentation. Ceci nous amène tout d'abord à examiner l'intérêt de la prise en compte des relations sémantiques entre les mots pour affiner la détection des points de passage d'un sujet à un autre. Nous proposons également un nouveau cadre théorique pour combiner différentes sources d'information sur les marqueurs linguistiques et les indices acoustiques, et employons une technique habituellement utilisée pour détecter les fins de phrase dans une transcription, en vue de repérer les frontières entre les sections.

Malgré son principe simple, la segmentation thématique pose la délicate question de ce que représente un *thème*. Nous débutons donc ce chapitre par des éléments de réponse pour définir cette notion. Nous poursuivons en passant en revue les travaux existants pour segmenter thématiquement un texte, en particulier une transcription, avant d'exposer une première version de notre modèle, basée uniquement sur la cohésion lexicale. Nous indiquons ensuite le protocole que nous avons suivi pour tester la qualité de la segmentation obtenue automatiquement et donnons les premiers résultats acquis avec ce modèle initial. Les deux sections suivantes sont consacrées à l'apport de connaissances supplémentaires pour améliorer la détection des ruptures thématiques, en considérant tout d'abord la prise en compte de relations sémantiques entre les mots puis celle des indices acoustiques et des marqueurs linguistiques. Nous discutons enfin de la qualité de la segmentation automatique obtenue, en suggérant des évolutions pour qu'elle soit intégrée plus efficacement dans notre chaîne de traitements.

7.1 Qu'est-ce qu'un thème ?

La notion de *thème* peut être décrite de manière intuitive comme ce dont on « parle » ; elle est cependant difficile à définir de manière consensuelle. En linguistique, le terme de thème désigne deux notions assez différentes. Une première, définie au niveau de la phrase, désigne un élément d'un énoncé qui est supposé connu par les participants à la communication, par opposition au rhème qui est l'information nouvelle apportée par l'énoncé [Ghadessy, 1995]. Ainsi, dans l'exemple « *Mon patron m'a donné une augmentation. Mais elle n'est pas énorme.* », on peut considérer que dans la première proposition « *mon patron* » est le thème et le reste le rhème mais qu'« *une augmentation* » passe du statut de constituant du rhème à celui de thème dans la phrase qui suit. Une seconde notion représente quant à elle le sujet dont traite un texte, une conversation ou une partie de l'un ou de l'autre. Il s'agit alors d'un élément qui intervient à plusieurs reprises dans une section de plusieurs phrases et c'est cette vision des choses à laquelle nous nous intéressons ici[1].

Brown et Yule suggèrent de définir le thème dans ce cadre en adoptant une méthodologie de type différentiel, *i.e.*, en caractérisant les changements de sujets plutôt que les thèmes eux-mêmes [Brown et Yule, 1983]. Ce principe conduit à découper le discours en une suite de segments, homogènes du point de vue du sujet abordé ; c'est une voie qui a été souvent suivie en TAL (*cf.* section 7.2.2.1). Les travaux de Rastier quant à eux s'appuient sur la notion d'isotopie appliquée au niveau intra-textuel et inter-textuel, celle-ci étant définie comme un effet de la récurrence au sein d'un texte d'une ou plusieurs unités sémantiques élémentaires, appelées sèmes [Rastier, 1995]. Selon ce point de vue, c'est le repérage de la répétition de ces sèmes qui permet ainsi de percevoir la cohérence d'une partie d'un document, et de ce fait d'identifier les segments théma-

[1] Afin de ne pas confondre avec l'opposition entre thème et rhème faite au niveau de la phrase, certains auteurs préfèrent employer le terme de « sous-thème » plutôt que de « thème » dans le cadre de la segmentation thématique [Hearst, 1997]. Nous choisissons quant à nous d'utiliser indifféremment les termes de sujet ou de thème pour désigner ce dont « parle » une section.

tiques. Il autorise en outre la caractérisation du sujet abordé au moyen d'une liste de sèmes [Rossignol, 2005].

Ces définitions de la notion de thème apportent des éléments pour segmenter un texte, en examinant les points de rupture et les répétitions de traits sémantiques. Elles demeurent toutefois imprécises quant à la granularité d'un thème, un texte pouvant aborder un sujet général qui peut être décomposé en plusieurs points plus précis. Si la caractérisation de ce qui doit figurer comme un thème ou non dans un document contient une part de subjectivité, elle s'appuie toutefois sur des conceptions communes entre les individus. Une expérience analysant les segmentations thématiques effectuées par différentes personnes sur 20 monologues montre ainsi un accord entre les annotateurs statistiquement significatif [Passonneau et Litman, 1993]. Pour notre part, nous cherchons à obtenir des thèmes assez généraux pour ne pas avoir à analyser de nombreuses sections limitées à quelques groupes de souffle pour un même document, mais toutefois assez spécifiques pour obtenir un corpus d'adaptation qui soit beaucoup plus pertinent qu'un corpus général sur une section donnée. Le thème « actualité internationale » est ainsi considéré comme trop imprécis, sauf si par exemple une émission sur le sport ne lui consacre que quelques secondes. Au contraire, la détection de chaque segment portant sur un point particulier de la guerre en Irak, tels que « le rôle de l'ONU », « la traque de Sadam Hussein » ou « la reconstruction de l'Irak », est jugée trop précise dans la mesure où un corpus qui traite de cette guerre en général semble déjà beaucoup plus pertinent que celui utilisé par défaut. On voit donc que suivant l'interprétation de ce que doit représenter un thème, le découpage souhaité est différent. Cette tâche de segmentation est d'autant plus difficile que mise à part l'existence de sujets très généraux, la description des sujets abordés ne peut être que contextuelle, ceux-ci étant souvent traités sous un angle et avec un niveau de développement qui varient d'un document à un autre.

Avant de présenter la méthode de segmentation que nous avons employée dans notre étude, nous exposons dans la section suivante les principales techniques développées pour partitionner un texte suivant les thèmes qu'il aborde.

7.2 Travaux existants

La segmentation thématique consiste à localiser dans un texte les points de changement de thème, de manière à obtenir des unités thématiquement homogènes. Cette opération, souvent étudiée en TAL, peut être utile dans différents cadres applicatifs. En RI, elle permet ainsi de produire des réponses plus pertinentes en ne retournant pas le document entier mais les portions adéquates [Salton *et al.*, 1996]. Elle peut en outre servir à structurer une large base de données en partitionnant chaque document en entités homogènes, puis en recherchant les relations entre ces unités pour construire des interfaces de navigation plus intuitives [Ide *et al.*, 2003]. Elle aide également à la construction de résumés automatiques [Reynar, 1998] et à la caractérisation des thèmes traités dans un corpus en extrayant une signature lexicale [Ferret et Grau, 2001]. Si, comme nous l'avons dit, la plupart des travaux s'intéressant à cette problématique sont

issus du domaine du TAL, notons cependant que les études qui se rapprochent le plus de notre cadre applicatif ont été réalisées lors des campagnes TDT (pour *Topic Detection and Tracking*) organisées par l'agence DARPA.

Les tâches de TDT ont pour but d'examiner les performances atteintes pour trouver et suivre des événements de manière automatique, parmi un ensemble d'émissions d'information [Allan *et al.*, 1998] ; l'une de ces tâches concernait plus spécifiquement la segmentation thématique. Le corpus traité incluait des dépêches de l'agence Reuters ainsi que des transcriptions manuelles de journaux diffusés sur CNN, seule une petite partie du corpus de CNN comptant 1000 unités thématiques ayant été transcrite automatiquement. Contrairement aux travaux qui ont été menés dans ce cadre, nous ne disposons que d'un volume beaucoup plus réduit (*cf.* section 3.1) puisque la totalité du corpus TDT couvrait 16 000 sections collectées sur une année. La segmentation de transcriptions automatiques demeure au final peu étudiée, notamment du fait de la présence de mots mal reconnus et de leurs spécificités par rapport aux textes du domaine de l'écrit.

Nous présentons dans ce qui suit les indices sur lesquels s'appuient les méthodes de segmentation, avant de décrire les principales techniques qui ont déjà été appliquées sur des textes.

7.2.1 Indices utilisés

Parmi les différentes sources d'information qui permettent de segmenter thématiquement un document, le critère de *cohésion lexicale* est le plus étudié. La notion de cohésion fait référence aux relations grammaticales et lexicales qui existent à l'intérieur d'un texte, en lui donnant une certaine unité et un sens. Parmi ces relations, on retrouve l'utilisation de pronoms qui font référence à un élément étant intervenu précédemment, la substitution d'un groupe de mots par un autre pour désigner la même chose, ou encore l'emploi de mots de liaison comme par exemple « *puis* », « *de plus* » ou « *parce que* » [Halliday et Hasan, 1976]. La cohésion lexicale fait plus particulièrement référence au fait que de nombreux mots d'une section thématique entretiennent des relations sémantiques. Cela se manifeste par la répétition de la même lexie, de lexies du même paradigme, telles que « BUS » et « VOITURE » qui désignent tous les deux des moyens de locomotion, ou encore par celle de collocations qui représentent des associations prédisposées entre deux mots, telles que « *peau* » et « *orange* ».

En pratique, l'emploi de la cohésion lexicale pour détecter automatiquement des ruptures thématiques s'appuie essentiellement sur le repérage de la répétition de mots [Reynar, 1994 ; Hearst, 1997 ; Heinonen, 1998 ; Choi, 2000 ; Utiyama et Isahara, 2001 ; Kehagias *et al.*, 2003 ; Ide *et al.*, 2003 ; Ji et Zha, 2003]. Du fait des difficultés rencontrées pour désambiguïser automatiquement chaque mot selon son sens, les méthodes de segmentation automatiques s'appuient sur l'étude des occurrences d'un ensemble de mots plutôt que d'un seul, ce qui limite les risques de considérer comme équivalents des emplois différents d'un même terme. Certains mots sont de plus parfois pondérés par un critère de significativité, en donnant un poids plus important aux termes qui apparaissent peu fréquemment dans un corpus de référence et qui sont susceptibles

d'être plus informatifs sur le thème traité que d'autres [Richmond *et al.*, 1997 ; Yaari, 1997 ; Ferret et Grau, 2001 ; Georgescul *et al.*, 2006b]. Ce principe semble toutefois moins primordial pour la segmentation qu'il ne l'est pour d'autres domaines comme la RI [Hearst, 1997]. Il est en revanche préférable de se restreindre aux mots lexicaux et d'éliminer certains termes d'emploi très courant qui contribuent peu au sens du texte, comme « *avoir* » et « *être* » [Reynar, 1994].

L'étude de la seule stricte répétition des mots ignore de nombreuses relations sémantiques comme la synonymie[2], l'hyperonymie[3], l'antonymie[4] ou plus généralement le partage de traits sémantiques. Pour remédier à ce problème, certaines techniques de segmentation recourent à des dictionnaires ou des thésaurus [Morris et Hirst, 1991 ; Kozima, 1993 ; Okumura et Honda, 1994 ; Jobbins et Evett, 1998] pour évaluer la proximité sémantique entre les mots. Ferret et Grau [2001] utilisent quant à eux un réseau de cooccurrences établi sur un corpus de grande taille pour étendre les mots du texte à segmenter par d'autres mots qui leur sont liés. Une autre façon de procéder consiste à apprendre la similarité lexicale en projetant chaque mot dans un espace qualifié de sémantique, en fonction de sa cooccurrence avec les autres mots [Kaufmann, 1999 ; Choi *et al.*, 2001 ; Bestgen, 2006] ; l'analyse sémantique latente (ASL) est une méthode qui est souvent utilisée dans ce cadre. Cette proximité peut également être obtenue sur un corpus de manière automatique en examinant la cooccurrence des mots au sein d'un même paragraphe ou d'un même document, les termes ayant tendance à apparaître ensemble souvent dans des unités thématiquement cohérentes pouvant être considérés comme fortement liées au niveau du sens [Rossignol, 2005 ; Caillet *et al.*, 2004]. L'utilisation de connaissances supplémentaires au critère de répétition ne conduit pas toujours à une amélioration de la qualité de la segmentation, notamment à cause de la difficulté à les acquérir et les employer [Kan *et al.*, 1998 ; Stokes *et al.*, 2004] ; l'ASL semble cependant apporter un gain [Choi *et al.*, 2001 ; Bestgen, 2006].

D'autres critères que celui de la cohésion d'un segment au niveau du vocabulaire peuvent être utilisés pour détecter les ruptures thématiques. Il est ainsi possible de recourir au repérage de marqueurs linguistiques comme les particules du discours « *ok* », « *right* » en anglais, ou des adverbes temporaux tels que « *maintenant* » [Passonneau et Litman, 1997 ; Reynar, 1999 ; Boufaden *et al.*, 2002]. Ces mots peuvent notamment être obtenus à partir d'un corpus segmenté en détectant les mots-clés ou les N-grammes qui apparaissent au voisinage des frontières thématiques [Reynar, 1998 ; Beeferman *et al.*, 1999 ; Galley *et al.*, 2003].

Dans certaines applications, le texte n'est qu'un média parmi d'autres et des indices différents peuvent être tout aussi informatifs. Dans le cas de flux de télévision par exemple, le changement de plan et notamment l'apparition du plateau télé dans un

[2]Relation d'équivalence entre deux lexies.

[3]Relation sémantique hiérarchique d'une lexie à une autre selon laquelle l'extension du premier terme, plus général, engloble l'extension du second, plus spécifique. « MAMMIFÈRE » est par exemple est un hyperonyme de « LION ».

[4]Relation entre deux lexies dont les sens se distinguent par la négation ou plus généralement par la mise en opposition d'une de leurs composantes. « CHAUD » et « FROID » sont ainsi deux antonymes.

journal sont des indicateurs importants de changement de thème. En ce qui concerne les documents sonores, plusieurs caractéristiques du signal ont déjà été employées pour la segmentation. La durée des pauses silencieuses est ainsi souvent utilisée [Passonneau et Litman, 1997 ; Dharanipragada *et al.*, 1999 ; Tür *et al.*, 2001], de même que la détection d'alternance de locuteurs [Tür *et al.*, 2001 ; Chen *et al.*, 2001 ; Galley *et al.*, 2003]. Les indices relevant de la prosodie ont été comparativement peu étudiés [Passonneau et Litman, 1997 ; Tür *et al.*, 2001] principalement parce qu'il est difficile de les extraire automatiquement. Pour [Passonneau et Litman, 1997], ils sont ainsi obtenus manuellement, tandis que Tür *et al.* [2001] se limitent à la variation de la fréquence fondamentale du signal. Une étude de l'emploi de l'ensemble de ces informations acoustiques pour segmenter des émissions d'information indique que celles concernant la durée des pauses silencieuses et dans une moindre mesure le changement de locuteur sont les plus pertinentes pour cette tâche [Tür *et al.*, 2001].

7.2.2 Types de méthodes de segmentation

Nous donnons ici un aperçu des différentes méthodes de segmentation, en distinguant celles qui n'utilisent que la cohésion lexicale, celles qui supposent l'existence de thèmes prédéfinis et celles qui combinent différents types de sources d'information.

7.2.2.1 Cohésion lexicale seule

Les techniques utilisant la cohésion lexicale s'appuient sur l'analyse de la distribution des mots ; une rupture thématique est alors établie dès que le vocabulaire employé dans une partie du texte subit des variations significatives. Pour ce faire, ces méthodes effectuent un calcul de similarité lexicale entre des parties adjacentes d'un même texte. Un point délicat à résoudre concerne le choix des unités élémentaires utilisées dans la comparaison de la distribution des mots. Dans le cas où le texte se présente sous la forme de paragraphes, établis par l'auteur lui-même et dont on suppose une certaine cohérence du point de vue des idées abordées, ces unités peuvent être employées pour l'apparition des thèmes [Yaari, 1997]. Malheureusement, tous les corpus ne sont pas découpés de cette manière, ce qui est notamment le cas du corpus oral que nous étudions. Des expériences menées par Tür *et al.* [2001] suggèrent, au moins pour la méthode qu'ils emploient, que le recours aux groupes de souffle donne des résultats similaires à l'usage des phrases pour la segmentation thématique de transcriptions automatiques.

Une autre manière de procéder consiste à construire des fenêtres de taille fixe, ce qui permet de disposer d'ensembles de grandeurs comparables [Hearst, 1997]. Nombre de méthodes basées sur la cohésion lexicale dérivent des techniques employées par TextTiling [Hearst, 1997]. La segmentation se traduit alors par l'exécution en séquence d'opérations (Fig. 7.1). Une première étape consiste à normaliser le texte en conservant uniquement les mots lexicaux les plus fréquents et en les ramenant à une forme plus simple grâce à des lemmatiseurs ou des *stemmers*. Cette opération est presque systématiquement pratiquée, quelle que soit la méthode de segmentation envisagée, car elle permet de réduire l'espace des mots à comparer, ce qui facilite les calculs de simila-

rité, tout en ne perdant que peu d'informations. L'étape suivante consiste à calculer un score de segmentation en construisant des fenêtres d'analyse comportant un nombre fixe de mots [Kozima, 1993 ; Hearst, 1997 ; Kaufmann, 1999 ; Ferret et Grau, 2001], de phrases [Richmond *et al.*, 1997 ; Jobbins et Evett, 1998 ; Ide *et al.*, 2003], voire de paragraphes [Heinonen, 1998]. La détection des points de rupture thématique est alors basée sur l'étude de la similarité entre une fenêtre gauche et une fenêtre droite de tailles identiques, que l'on déplace selon un pas donné de façon à parcourir l'ensemble du texte.

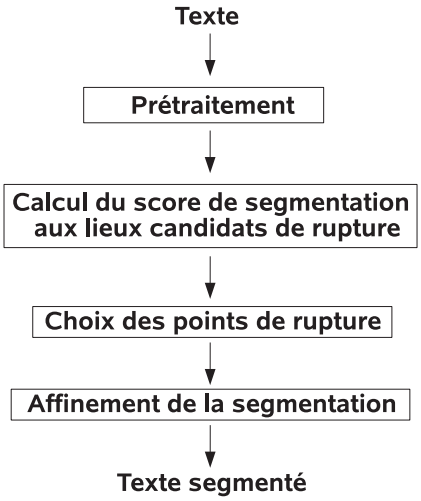

FIG. 7.1: *Principe général de la segmentation basée sur la cohésion lexicale seule*

TEXTTILING emploie deux méthodes principales de calcul de score. La première consiste à représenter le contenu de chaque fenêtre sous la forme d'un « sac de mots », formalisé par un vecteur dont les valeurs représentent le nombre d'occurrences de chaque mot (ou plus exactement de chaque lemme). Tous les 20 mots par exemple, l'algorithme calcule la ressemblance entre les contextes gauche et droit de ce point de focus, au moyen d'une distance par cosinus. La seconde méthode repose sur la notion de *chaîne lexicale* proposée par [Morris et Hirst, 1991] et consistant à repérer les occurrences rapprochées d'un même terme. Une fois les scores établis à différentes positions du texte, l'étape suivante de TEXTTILING consiste à repérer les minima de la similarité mesurée entre des fenêtres adjacentes (Fig. 7.2). Les baisses jugées les plus significatives sont alors considérées comme des ruptures thématiques, puis parfois ramenées à la frontière de phrase la plus proche. La segmentation ainsi obtenue peut être affinée dans une dernière étape en regroupant des segments adjacents jugés trop similaires [Dharanipragada *et al.*, 1999 ; Ide *et al.*, 2003] ou en déplaçant les frontières de segments thématiques d'après le repérage de marqueurs linguistiques [Hurault-Plantet *et al.*, 2006].

Il existe bien d'autres méthodes que celles employées par TEXTTILING. Une variante du calcul de similarité lexicale entre fenêtres repose sur une représentation graphique

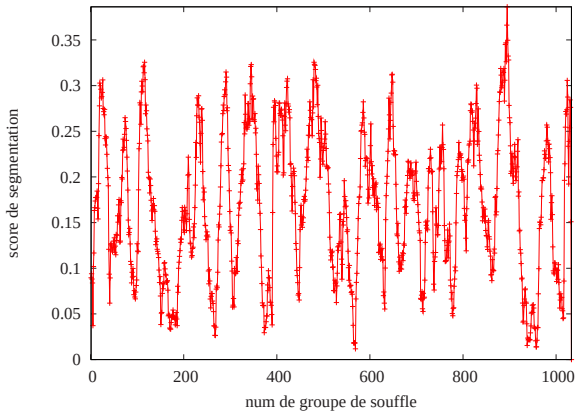

FIG. 7.2: *Évolution de la mesure de similarité par cosinus sur une transcription automatique d'une émission d'1 heure extraite du corpus de développement d'*ESTER*, en utilisant des fenêtres de* 25 *groupes de souffle*

de la répétition de termes [Reynar, 1994; 1998 ; Choi *et al.*, 2001 ; Ji et Zha, 2003]. Une méthode particulièrement originale, modélisant le problème de la segmentation thématique par la recherche de celle qui conduit aux segments les plus homogènes du point de vue lexical, est proposée par Utiyama et Isahara [2001]. Ce segmenteur est intéressant sur plusieurs points. Il est tout d'abord exprimé dans un cadre probabiliste, ce qui facilite son utilisation conjointe avec d'autres sources d'information, notamment celles relevant de l'acoustique. Il est donné en outre sous une forme qui ne demande pas, selon ses auteurs, de paramétrage, au contraire par exemple de TEXTTILING qui nécessite de fixer la taille des fenêtres d'analyse, le pas de déplacement de ces fenêtres, un seuil de détection des minima locaux et éventuellement une méthode de lissage des scores. Plusieurs évaluations comparant la méthode probabiliste avec d'autres montrent également qu'elle se situe toujours parmi les meilleures [Choi *et al.*, 2001 ; Georgescul *et al.*, 2006a]. Pour toutes ces raisons, c'est elle que nous avons retenue dans notre système de segmentation. Nous la décrivons plus en détail dans la section 7.3.2.

7.2.2.2 Modèles thématiques

Certaines méthodes s'appuient sur la détection de thèmes prédéfinis pour segmenter un document. Elles présentent l'avantage d'identifier le thème traité par chaque section mais soulèvent le problème de la construction de la liste des sujets abordés. Une première solution consiste à définir manuellement un ensemble de thèmes très généraux, qui couvrent la majorité des documents à traiter, au risque toutefois d'avoir une granularité insuffisante pour détecter convenablement les ruptures thématiques [Bigi *et al.*, 1998 ; Brun, 2003]. Une autre façon de faire est d'extraire automatiquement un nombre défini de sujets abordés dans un corpus, en rassemblant dans un même groupe les sections similaires [Van Mulbregt *et al.*, 1999 ; Li et Yamanishi, 2000 ; Brants, 2000 ; Blei et

Moreno, 2001 ; Akita *et al.*, 2007 ; Sako *et al.*, 2007]. Une technique très employée pour utiliser la détection de thème au cours de la segmentation est celle proposée par [Van Mulbregt *et al.*, 1999] ; initialement élaborée pour la campagne TDT, elle a inspiré beaucoup d'autres méthodes par la suite. Son principe est de modéliser le problème de la segmentation par un MMC, dont les observations sont des phrases et les variables cachées des thèmes. À l'heure actuelle, c'est une autre méthode probabiliste basée sur l'analyse sémantique latente probabiliste qui est plutôt employée [Brants, 2000 ; Akita *et al.*, 2007 ; Sako *et al.*, 2007].

7.2.2.3 Méthodes multi-sources

Les techniques généralement utilisées pour combiner différentes sources d'information, aussi bien lexicales qu'acoustiques, sont les arbres de décision [Passonneau et Litman, 1997 ; Dharanipragada *et al.*, 1999 ; Tür *et al.*, 2001 ; Galley *et al.*, 2003], les modèles à maximum d'entropie [Beeferman *et al.*, 1999 ; Reynar, 1999], les modèles SVM [Georgescul *et al.*, 2007] et les MMC [Tür *et al.*, 2001]. Cette dernière méthode, dont nous nous inspirons par la suite (*cf.* section 7.6), recourt d'une part au modèle thématique proposé par [Van Mulbregt *et al.*, 1999] pour modéliser les thèmes traités par chaque groupe de souffle W_i, et d'autre part, aux arbres de décision pour utiliser les caractéristiques acoustiques F_i établies au voisinage des fins de groupes de souffle W_i. Parmi les indices utilisés F_i, figurent la durée de la pause silencieuse entre W_{i-1} et W_i ou la différence entre la fréquence fondamentale moyenne du dernier mot de W_{i-1} et celle du premier mot de W_i[5]. Les séquences d'états du MMC ont pour forme $Q = (t_1, b_1 \dots t_i, b_i \dots t_\ell, b_\ell)$ où t_i représente le thème associé à W_i et b_i le fait qu'un changement de thème intervient ou non entre W_{i-1} et W_i. Le document sonore est ainsi décrit par ses informations textuelles $W = W_1^\ell$ et acoustiques $F = F_1^\ell$. Les probabilités sont établies par le MMC de la manière suivante :

$$P(W, F|Q) = \prod_{i=1}^{\ell} P(W_i|t_i) \prod_{i=1}^{\ell} P(F_i|b_i, W) \tag{7.1}$$

où $P(W_i|t_i)$ est calculé par un modèle thématique lexicale et $P(F_i|b_i, W)$ est obtenu en modélisant la relation entre la présence d'une frontière thématique et les caractéristiques acoustiques F_i. Suivant les expériences menées par Tür *et al.* [2001], ce mode de calcul s'avère plus performant qu'un arbre de décision pour combiner l'ensemble des sources d'information.

Durant l'aperçu des méthodes de segmentation existantes que nous avons donné, nous avons distingué les modèles thématiques des autres techniques basées sur la cohésion lexicale. Au cours de nos travaux, nous avons choisi de ne pas nous baser sur l'existence de thèmes prédéfinis qui, à moins de rester très généraux, sont impossibles à énumérer de manière exhaustive. Nous avons plutôt préféré employer une méthode de segmentation

[5]Cette mesure de la différence de fréquences fondamentales permet de mesurer un changement qui se produit sur le plan de la prosodie entre deux groupes de souffle.

qui suppose peu de connaissances *a priori* et qui est basée sur une approche probabiliste qui a déjà fait ses preuves pour intégrer différents types d'information obtenus à partir du document traité. Nous présentons dans les sections suivantes notre démarche pour segmenter une transcription automatique, en débutant par la description du modèle basé uniquement sur la cohésion lexicale.

7.3 Modèle de cohésion lexicale

Dans un premier temps, nous considérons la transcription à segmenter comme du texte, en ignorant les informations acoustiques que contient le document. La méthode retenue, proposée par Utiyama et Isahara [2001], requiert en entrée une représentation sous la forme d'un « sac de mots » de chaque phrase du texte étudié. Étant donné que nous opérons sur des transcriptions automatiques, nous considérons plutôt un découpage en groupes de souffle. Si le principe de la technique reste le même, cette façon de procéder conduit toutefois à examiner des unités plus courtes que des phrases. Nous exposons dans cette section les traitements que nous appliquons au préalable sur le document à traiter, avant de présenter en détail le modèle que nous reprenons, puis les modifications que nous y apportons.

7.3.1 Prétraitement

Afin de ne retenir que les mots susceptibles d'être informatifs pour la segmentation thématique et de les ramener à une forme simple, nous étiquetons la transcription à l'aide de notre étiqueteur morpho-syntaxique, puis nous la lemmatisons à l'aide de FLEMM [Namer, 2000]. Ces deux opérations, habituelles avant de procéder à l'application de méthodes basées sur la cohésion lexicale, nous permettent de ne retenir que les noms, les verbes et les adjectifs, *i.e.*, les classes porteuses de sens. Nous éliminons également les verbes « *avoir* », « *être* », « *falloir* », « *devoir* », « *pouvoir* » et « *vouloir* » qui, employés la plupart du temps en tant que verbes modaux et auxiliaires, jouent un moindre rôle que les autres verbes du point de vue des informations sémantiques. La suppression des deux premiers, « *avoir* » et « *être* », est particulièrement importante pour les transcriptions automatiques, étant donné qu'ils sont souvent insérés ou omis à tort par le système de RAP. Une fois le texte ainsi traité, nous lui appliquons la méthode de cohésion lexicale que nous décrivons ci-après.

7.3.2 Modèle initial

Avant de proposer des évolutions de la méthode originale, nous exposons ici le modèle de cohésion lexicale tel que conçu par Utiyama et Isahara [2001]. En résumé, cette technique suit une approche probabiliste recourant à la loi de Laplace pour modéliser la cohérence des sections thématiques, représentées comme des « sacs de mots ». Le problème de segmentation d'un texte $W = w_1^n$ de n mots (ou plus exactement de n lemmes correspondant à des mots lexicaux) en plusieurs sections $S = S_1^m$ est alors

défini par la recherche de la segmentation $\hat{S}$ la plus probable :

$$\hat{S} = \arg\max_{S} P(S|W) \ , \tag{7.2}$$

ce qui, de manière similaire à la modélisation de la RAP selon un critère MAP, se réécrit par :

$$\hat{S} = \arg\max_{S} P(W|S)P(S) \ . \tag{7.3}$$

7.3.2.1 Définition de $P(W|S)$

L'évaluation de $P(W|S)$ est réalisée en notant les n_i mots contenus dans un segment S_i par $W_i = w_1^{(i)} w_2^{(i)} \dots w_{n_i}^{(i)}$. Le texte W peut alors être représenté sous la forme $W = W_1^m$ et $P(W|S)$ devient :

$$P(W|S) = P(W_1 \dots W_m|S) \ . \tag{7.4}$$

En supposant que chaque section thématique forme une unité indépendante du restant du texte, on a :

$$P(W|S) = \prod_{i=1}^{m} P(W_i|S) = \prod_{i=1}^{m} P(W_i|S_i) \ . \tag{7.5}$$

En outre, chaque section étant modélisée par un « sac de mots », l'ordre des mots n'a pas d'influence sur le calcul des probabilités, ce qui conduit à écrire :

$$P(W|S) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} P(w_j^{(i)}|S_i) \ . \tag{7.6}$$

Les valeurs $P(w_j^{(i)}|S_i)$ sont ensuite modélisées par la loi de Laplace qui permet de prédire l'apparition d'un événement à valeurs discrètes, en ayant observé un certain nombre de fois cet événement. Dans notre contexte, en considérant que l'on a observé les mots W_i du segment S_i, la probabilité de voir apparaître à nouveau un mot $w_j^{(i)}$ dans ce segment est exprimée par :

$$P(w_j^{(i)}|S_i) = \frac{f_i(w_j^{(i)}) + 1}{n_i + k} \tag{7.7}$$

où $f_i(w_j^{(i)})$ donne le nombre d'occurrences de $w_j^{(i)}$ dans S_i et k le nombre de mots différents dans le texte à segmenter. D'un point de vue intuitif, cette probabilité favorise les segments homogènes du point de vue lexical car elle augmente quand les mots apparaissent plusieurs fois et diminue si beaucoup de mots sont différents.

7.3.2.2 Définition de $P(S)$

La probabilité $P(S)$ mesure directement le nombre de segments souhaités dans un texte, ce qui dépend fortement de la nature du document à traiter et de la granularité du découpage souhaité. Les auteurs de cette méthode propose une modélisation *a priori*

de cette valeur en recourant à la théorie de l'information. $P(S)$ est alors obtenue à partir de la description de la longueur $l(S)$ de S, *i.e.*, du nombre minimum d'informations à envoyer pour que la segmentation soit déterminée, ce qui s'exprime par :

$$P(S) = 2^{-l(S)} = n^{-m} \tag{7.8}$$

puisque $l(S) = m \log_2 n$ bits.

7.3.2.3 Méthode d'implémentation

D'après l'équation (7.3), la recherche de la meilleure segmentation revient à énumérer tous les découpages possibles. Afin de restreindre l'espace des hypothèses, des contraintes sont imposées. Les auteurs de la méthode proposent ainsi de considérer que les ruptures thématiques ne puissent intervenir qu'à des fins de phrase et suggèrent de recourir à la programmation dynamique afin d'augmenter la rapidité des calculs. En pratique, le coût à optimiser est la suivante :

$$C(S) = -\log P(W|S)P(S) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} - \left[\log \frac{f_i(w_j^{(i)}) + 1}{n_i + k}\right] + m \log n \ . \tag{7.9}$$

En notant par $C(W_i|n,k)$ le coût d'un segment W_i :

$$C(W_i|n,k) = \sum_{j=1}^{n_i} \left[\log \frac{n_i + k}{f_i(w_j^{(i)}) + 1}\right] + \log n \ , \tag{7.10}$$

le problème de segmentation se réécrit :

$$\hat{S} = \arg\min_S C(S) = \arg\min_S \sum_{i=i}^{m} C(W_i|n,k) \ . \tag{7.11}$$

Soit ℓ le nombre de groupes de souffle du document étudié. Le coût de chaque segmentation $C(S)$ est évalué après avoir construit un graphe $G = (V, E)$, où $V = \{g_i | 0 \leq i \leq \ell\}$ correspond aux sommets et $E = \{e_{ij} | 0 \leq i < j \leq \ell\}$ aux arcs (Fig 7.3). g_i désigne la position entre la phrase s_i et s_{i+1}, tandis que e_{ij} représente le segment constitué des phrases $s_{i+1} \ldots s_j$ et de coût $C(s_{i+1} \ldots s_j|n,k)$. La recherche de la meilleure segmentation peut alors se faire en deux étapes. Le graphe G est tout d'abord construit pour l'ensemble du texte, en évaluant le coût de chacun des arcs. Trouver les points de rupture thématique revient ensuite à trouver le chemin de coût minimal entre g_0 et g_ℓ en utilisant un algorithme de programmation dynamique.

7.3.3 Modification du modèle original

Nous proposons dans cette section plusieurs évolutions du modèle précédemment décrit en fonction des contraintes posées par le type de documents étudiés. Nous faisons tout d'abord l'hypothèse que le passage d'un thème à un autre ne peut se produire qu'à

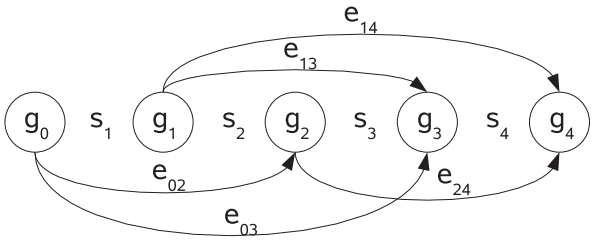

FIG. 7.3: *Exemple de graphe construit sur un texte de 4 phrases, en restreignant la taille des sections à 2 ou 3 phrases*

la fin d'un groupe de souffle. Ce postulat nous paraît tout aussi valable que celui reposant sur la segmentation en phrases pour les textes écrits puisque la reprise de son souffle ou la prise de parole d'un autre locuteur semblent être *a priori* des événements qui interviennent au début et à la fin d'un thème. Ceci est confirmé par l'analyse manuelle de 8 heures d'émissions d'information que nous avons pu effectuer et qui n'a montré que deux changements de sujet au cours d'un groupe de souffle. Au cours de nos expériences, nous avons de plus supposé que les sections thématiques ne pouvaient contenir qu'entre 2 et 200 groupes de souffle, ce qui restreint le nombre d'arcs du graphe à construire pour la recherche de la meilleure segmentation.

Nous apportons également des modifications au niveau des calculs de $P(W|S)$ et de $P(S)$. En ce qui concerne $P(W|S)$, la cohésion d'un segment est évaluée dans la version initiale en construisant un modèle de langage sur cette section, basé sur la loi de Laplace, puis en évaluant la capacité de ce modèle à prédire les mêmes données que celles qui ont servi à son apprentissage. Nous avons voulu étudier le comportement d'autres types de ML qui sont eux employés dans le cadre de la RAP. La taille de chaque section conduisant à un corpus de taille extrêmement réduite nous a amené à ne considérer que des ML ne prenant pas en compte le contexte d'apparition des mots dans le calcul des probabilités. La première loi que nous avons examinée est une simple probabilité unigramme, sans utiliser de lissage :

$$P(w_j^{(i)}|S_i) = \frac{f_i(w_j^{(i)})}{n_i} \quad . \tag{7.12}$$

La seconde est la loi naturelle de Ristad, dont il est montré expérimentalement qu'elle surpasse généralement la loi de Laplace pour prédire une suite d'événements discrets [Ristad, 1995]. Dans le modèle que nous utilisons, elle s'exprime de la manière suivante :

$$P(w_j^{(i)}|S_i) = \begin{cases} \frac{f_i(w_j^{(i)})}{n_i} & \text{si } q = k \\ \left(\frac{f_i(w_j^{(i)})}{n_i}\right) \frac{n_i(n_i+1)+q(1-q)}{n_i^2+n_i+2q} & \text{si } q < k \end{cases} \tag{7.13}$$

où q est le nombre de lemmes différents dans S_i et k le nombre de lemmes différents dans le texte à segmenter.

Le mode de calcul de $P(S)$ proposé par défaut permet de définir un ordre de grandeur sur la pénalité d'insertion d'une nouvelle frontière thématique. Suivant des expériences que nous avons menées, la valeur calculée est cependant à modifier suivant la longueur du document étudié et la taille moyenne des segments désirée. Nous introduisons donc un coefficient γ qui nous permet de faire évoluer la finesse du découpage obtenu. $P(S)$ devient alors :

$$P(S) = (n^{-m})^{\gamma} \ , \tag{7.14}$$

ce qui se traduit par la modification de $\log n$ en $\gamma \log n$, au niveau du calcul du coût $C(W_i|n,k)$ (*cf.* équation (7.10)).

La description du modèle de cohésion lexicale que nous employons étant faite, nous nous proposons dans la section suivante de l'évaluer.

7.4 Méthode d'évaluation

L'évaluation de la qualité de la segmentation pose deux problèmes principaux, qui sont la construction d'un découpage de référence d'un corpus de test et la définition d'une métrique. Nous présentons donc dans la suite de cette section nos choix sur ces deux points de difficulté, avant d'exposer les résultats obtenus au moyen de notre modèle de cohésion lexicale.

7.4.1 Corpus de référence

Compte tenu des différentes interprétations de ce que représente un thème et du fait que le passage d'un sujet à un autre peut être progressif, il est possible de construire plusieurs segmentations de référence pour un même document. De façon à limiter ce problème, il est d'ailleurs souvent recommandé d'avoir plusieurs annotateurs de façon à pouvoir dégager un consensus dans des cas litigieux [Passonneau et Litman, 1993]. Pour contourner ce problème, les corpus de référence sont souvent constitués de manière artificielle en concaténant des documents divers [Choi, 2000]. Toutefois, cette solution n'est pas totalement satisfaisante puisque dans la réalité, les changements de thèmes s'effectuent rarement de manière aussi brutale que le passage d'un texte à un autre.

Étant donnée la particularité de notre contexte applicatif, nous utilisons quant à nous le corpus ESTER sur lequel nous mesurons l'influence de notre approche sur le WER, pour évaluer notre méthode de segmentation. Ces données sont déjà divisées en entités journalistiques ; toutefois ce découpage n'est pas homogène, comme nous l'avons déjà mentionné en section 3.1, ce qui nous a conduit à produire nous-même une segmentation de référence sur les corpus de développement et de test. Nous nous sommes restreint à quatre émissions, les deux heures diffusées sur RFI possédant une segmentation manuelle en groupes de souffle de taille anormalement réduite suite aux choix de l'annotateur, et les émissions de RTM ayant un format très court de 10 à 20 minutes. Notre corpus d'étude se limite ainsi à quatre heures, pour lesquelles nous précisions dans le tableau 7.1 les performances obtenues par le système IRENE, auquel

		sans POS	$s'(w_1^n)$
dév.	France Inter - 18/04/03 - 7h-8h	18,0	17,5
	France Info - 18/04/03 - 12h-13h	18,6	17,8
test	France Inter - 18/04/03 - 8h-9h	17,3	15,9
	France Info - 18/04/03 - 17h-18h	19,0	18,1

TAB. 7.1: *Erreurs de transcription (%) mesurées sur les émissions étudiées pour l'adaptation thématique à l'aide de la mesure LER'_{lex} définie à la section 5.2.4.2*

nous adjoignons notre module prenant en compte les informations morpho-syntaxiques (*cf.* chapitre 5).

La comparaison des erreurs de transcription obtenues suivant un critère MAP, sans l'utilisation des POS ou avec, montre l'intérêt des informations morpho-syntaxiques pour réduire le nombre de mots mal reconnus (première et troisième colonnes). Les mêmes observations peuvent être faites en ce qui concerne les lemmes associés aux mots que nous retenons pour notre modèle de cohésion lexicale (deuxième et quatrième colonnes).

Si le recours à différents annotateurs est généralement souhaitable, nous avons constaté sur notre corpus, constitué d'émissions d'information, autrement dit de documents préparés et bien structurés, que les ruptures thématiques sont souvent bien marquées. Les points litigieux sont ainsi beaucoup moins nombreux que pour d'autres types de données, tels que les dialogues enregistrés lors de réunions. Nous avons cependant identifié deux cas particuliers à traiter. Premièrement, nous avons conservé une catégorie de sections, appelée *fillers*, déjà proposée par la segmentation initiale du corpus ESTER et correspondant à des groupes de souffle n'étant associés véritablement à aucun thème. On retrouve notamment parmi ces segments des indications sur la fréquence de la chaîne de radio ou sur l'heure, telles que « *France Info il est huit heures* », ou encore des formules de politesse, pour souhaiter la bienvenue aux auditeurs ou les remercier de leur écoute par exemple. Nous conservons bien entendu tous les mots figurant parmi les *fillers*, dans les transcriptions sur lesquelles nous appliquons nos méthodes de segmentation automatiques. En revanche, nous supprimons ces groupes de souffle lors de leur évaluation, ce qui permet de considérer comme identiques des frontières thématiques détectées juste avant ou juste après un *filler*.

Deuxièmement, il existe pour certains passages plusieurs façons de découper une émission suivant la granularité souhaitée des thèmes. Nous avons donc distingué deux niveaux de thèmes, l'un plus général et l'autre plus spécifique, qui dépendent de l'émission traitée et des liens entretenus entre les points abordés. Ainsi, pour une émission d'une heure abordant essentiellement les actualités internationales et nationales, nous définissons le thème « *Condolezza Rice* » comme étant général si le portrait fait sur cette personnalité est traité durant plusieurs minutes et n'est pas suivi ou précédé directement de sujets présentant des liens évidents avec lui. En revanche, le sujet du « *football* », pourtant général dans d'autres contextes, est considéré comme étant un sous-thème du

thème « *sport* » s'il fait partie d'un court segment d'une page sportive. Lors de notre évaluation, partant du principe qu'avoir des segments ne traitant que d'un thème assez général pouvait être suffisant pour améliorer la qualité de la transcription, nous n'avons considéré que les ruptures thématiques entre thèmes, ce qui a pour conséquence que des frontières entre sous-thèmes pourtant détectées correctement par la méthode de segmentation automatique sont assimilées comme fausses.

7.4.2 Métriques d'évaluation

Les métriques qui sont souvent retenues pour évaluer la qualité de la segmentation thématique sont les valeurs classiques de rappel et de précision :

$$\text{rappel}_0 = \frac{\text{nombre de frontières correctement détectées}}{\text{nombre de frontières de la segmentation de référence}} \tag{7.15}$$

$$\text{précision}_0 = \frac{\text{nombre de frontières correctement détectées}}{\text{nombre de frontières signalées}} \ . \tag{7.16}$$

Pour pallier le fait que les erreurs de segmentation à quelques mots près sont *a priori* sans trop de conséquence pour notre tâche d'adaptation, nous calculons des valeurs complémentaires $(\text{rappel}_1, \text{précision}_1)$ et $(\text{rappel}_2, \text{précision}_2)$ en considérant comme correctes des frontières établies respectivement avec au plus un ou deux groupes de souffle de différence par rapport à la segmentation de référence. Notons que plusieurs métriques ont été également proposées spécifiquement au problème de segmentation thématique pour pénaliser plus lourdement les frontières éloignées de celles de référence, en particulier P_k proposée par [Beeferman *et al.*, 1999] et *WindowDiff* [Pevzner et Hearst, 2002]. Néanmoins, ces métriques sont régulièrement remises en cause [Pevzner et Hearst, 2002 ; Georgescul *et al.*, 2006a] et ont été développées pour d'autres cadres applicatifs que les nôtres.

Étant donné notre contexte d'utilisation de la segmentation thématique, les mesures de qualité de rappel et de précision ne restent que des indicateurs du découpage produit. Notre objectif final étant d'obtenir une amélioration de la qualité de la transcription, les sections obtenues peuvent ne pas être identiques à celles produites manuellement pour peu qu'elles permettent de retourner des corpus d'adaptation pertinents pour modifier le ML général. Le critère qui semble le plus important est le rappel car il est préférable de détecter toutes les ruptures thématiques pour avoir des segments homogènes. Les sections obtenues ne doivent pas toutefois être trop petites. Nous mesurons donc la taille moyenne des segments produits, ainsi que la proportion de sections traitant d'un seul thème.

7.4.3 Premiers résultats

Le modèle de cohésion lexicale utilisé ne dépend que d'un paramètre γ, dont la valeur influe directement sur la taille moyenne des sections produites. Durant nos expérimentations, nous employons le corpus de développement pour étudier le comportement de ce

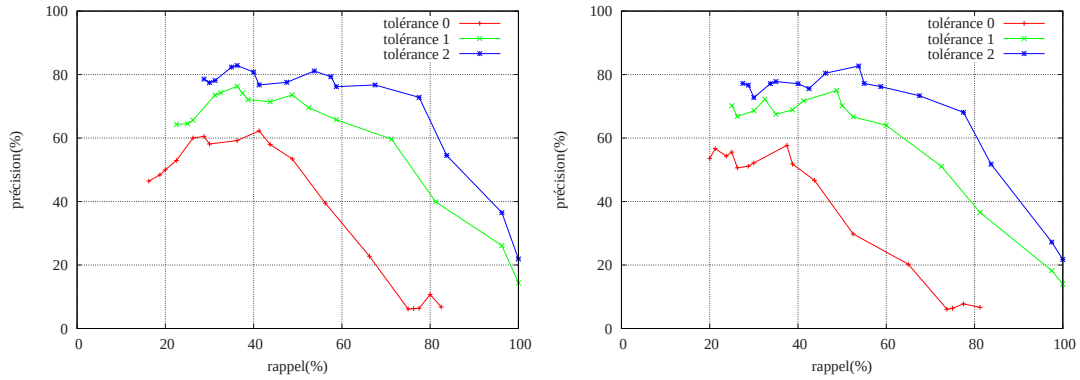

FIG. 7.4: *Évolution sur le corpus de développement de la précision en fonction du rappel pour une segmentation de la transcription manuelle (à gauche) et automatique (à droite)*

coefficient sur la qualité de la segmentation et le corpus de test pour vérifier la stabilité des résultats obtenus une fois γ fixé.

Les trois courbes affichées sur chacun des graphiques de la figure 7.4 indiquent les valeurs de précision et rappel calculées avec une tolérance d'erreur de 0, 1 et 2 groupes de souffle, en faisant varier γ sur le corpus de développement. La comparaison des résultats obtenus à partir de la transcription de référence (graphique de gauche) et celle produite par le système de RAP (graphique de droite) montre sans surprise de meilleurs résultats pour celle transcrite manuellement puisqu'elle conduit à rappel égal à de meilleures valeurs de précision. Les différences observées sont toutefois peu importantes, surtout pour les courbes réalisées avec une marge d'erreur de 1 ou 2 groupes de souffle. Ceci souligne le fait que les mots porteurs de sens sont globalement bien reconnus par le système de RAP.

Les valeurs de rappel égales à 100 % indiquées sur les graphiques correspondent à des sections limitées à deux groupes de souffle, ce qui revêt peu d'intérêt pour la segmentation. Au contraire, les fortes valeurs de précision consistent à ne diviser l'émission qu'en quelques sections. Nous fournissons sur la figure 7.5 un autre mode de présentation des résultats que nous venons d'exposer en donnant l'évolution du rappel en fonction de la taille moyenne des sections trouvées par la méthode automatique. Cette dernière valeur s'avère plus intuitive et c'est elle que nous utilisons pour fixer la valeur du coefficient γ en la comparant avec le nombre moyen de sections (égal à 23,3 groupes de souffle) du découpage manuel du corpus de développement.

À titre de comparaison, nous avons mené des expériences sur le même corpus avec la méthode de TEXTTILING basée sur la similarité par cosinus. Les résultats relevés avec cette technique sont nettement moins bons que ceux obtenus avec la méthode probabiliste puisque, pour des sections proposées de taille moyenne égale à 20 groupes de souffle, nous obtenons par exemple une précision inférieure à 35 %, alors que nous avions une valeur voisine de 50 % auparavant.

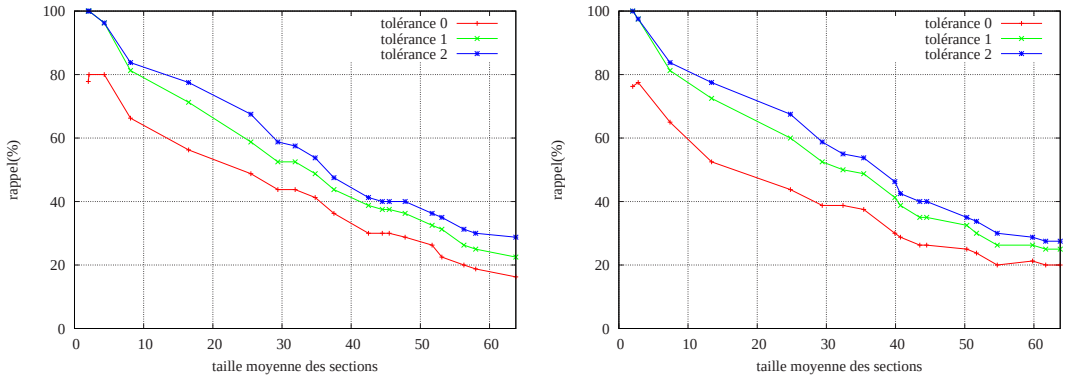

FIG. 7.5: *Évolution sur le corpus de développement du rappel en fonction de la taille moyenne des sections pour une segmentation de la transcription manuelle (à gauche) et automatique (à droite)*

taille moy.	rappel_0	précision_0	rappel_2	précision_2	sections homogènes
15,7	57,4	36,1	80,3	67,0	78/99
22,2	44,3	39,7	72,1	66,2	46/70

TAB. 7.2: *Qualité de la segmentation produite sur la transcription automatique des émissions du corpus de test en recourant à la loi de Laplace*

Afin de tester le comportement de la méthode probabiliste sur d'autres données, nous exposons dans le tableau 7.2 les valeurs de rappel et de précision calculées sur le corpus de test pour deux tailles moyennes de section (colonne 1), avec différents seuils de tolérance (colonnes 2 à 5). La dernière colonne indique le nombre de sections ne traitant que d'un seul thème parmi l'ensemble des segments trouvés. En fixant γ sur le corpus de développement, la taille moyenne des sections est restée sensiblement la même sur le corpus de test que celle trouvée initialement, ce qui montre une certaine stabilité de l'effet de ce coefficient sur la segmentation. Les résultats proposés, très similaires à ceux déjà constatés sur le corpus de développement, nous permettent d'obtenir un nombre relativement important de segments homogènes sur le plan thématique. Les valeurs de rappel restent cependant assez faibles. Tous les résultats que nous avons présentés ont été obtenus avec la loi de Laplace, proposée par défaut pour le calcul de $P(W|S)$, des expériences complémentaires menées avec les deux autres modes de calcul ayant conduit à des valeurs de précision et de rappel un peu moins bonnes. Le tableau 7.3 présente ainsi les résultats obtenus sur le corpus de test avec une loi de probabilité unigramme. La raison qui peut expliquer le manque de performance de la loi naturelle de Ristad est le fait que la plupart des mots lexicaux n'apparaissent qu'une fois localement, ce qui a pour conséquence que le nombre d'événements observés est proche de leur nombre total d'occurrences.

taille moy.	rappel_0	précision_0	rappel_2	précision_2	sections homogènes
13,0	36,1	18,7	63,9	32,2	83/120
20,8	41,0	34,3	59,0	49,3	47/75

TAB. 7.3: *Qualité de la segmentation produite sur la transcription automatique des émissions du corpus de test en recourant à la loi unigramme*

En règle générale, tout au moins sur notre corpus, les journalistes évitent de réemployer les mêmes termes en recourant à des synonymes. Afin de contourner cette difficulté de nature à gêner notre méthode de cohésion lexicale basée uniquement sur l'examen des occurrences de chaque mot, nous examinons dans la section suivante d'autres relations entre les mots que celles du simple phénomène de répétition.

7.5 Prise en compte d'un réseau de cooccurrences

L'examen des liens sémantiques entre les mots permet de se rendre compte de la richesse de la langue pour exprimer des choses similaires. Cela conduirait dans notre contexte à pouvoir identifier plus finement que jusqu'à présent l'homogénéité d'une section traitant un thème particulier. Afin d'étudier les potentialités de ces connaissances, nous employons une méthode d'acquisition assez simple basée sur l'analyse des cooccurrences. Dans cette section avant tout prospective, nous présentons les principes de la technique précédemment développée par Ferret [1998] pour obtenir les relations entre les mots, en proposant une manière pour inclure ces informations dans notre modèle. Nous indiquons ensuite les résultats expérimentaux obtenus, en discutant de l'intérêt de la méthode pour la segmentation thématique.

7.5.1 Méthode

Nous suivons la procédure proposée par Ferret [1998] pour découvrir les relations entre les mots qui ont tendance à apparaître souvent ensemble. Pour ce faire, la similarité entre deux termes x et y est calculée en examinant leur cooccurrence au moyen de l'information mutuelle $I(x, y)$:

$$I(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \tag{7.17}$$

où $P(i)$ s'obtient à partir de la fréquence d'apparition $f(i)$ de i mesurée sur un corpus, et $P(x, y)$ à partir du nombre de fois $f(x, y)$ où x et y apparaissent ensemble. $f(x, y)$ est déterminé en déplaçant une fenêtre d'analyse sur le corpus et en considérant que les occurrences de deux termes sont voisines si elles se produisent avec une distance inférieure à 20 mots. Si T_c est la taille du corpus utilisé pour ces calculs, l'information mutuelle s'obtient finalement par :

$$I(x, y) = \log \left(T_c \frac{f(x, y)}{f(x)f(y)} \right) \quad . \tag{7.18}$$

Afin d'obtenir une valeur comprise entre 0 et 1, cette valeur est normalisée par un maximum $I_{\max}$:

$$coh(x,y) = \frac{I(x,y)}{I_{\max}} \tag{7.19}$$

avec :

$$I_{\max} = \log[T_c^2(T_f) - 1] \tag{7.20}$$

en remarquant que $f(x,y) \leq T_c(T_f - 1)$ où T_f est la taille de fenêtre, $f(x) \geq 1$ et $f(y) \geq 1$.

En ne retenant que les liens dont les valeurs $coh(x,y)$ dépassent un certain seuil, nous obtenons un réseau de relations. Certaines de celles-ci sont peu évidentes sur le plan sémantique et sont dues à l'apparence fortuite du couple de mots dans le même voisinage ; le fait de fixer un seuil élevé pour $f(x,y)$ permet de limiter ce bruit. Nous avons ainsi conservé uniquement les relations telles que $I(x,y) \geq 0,15$ et $f(x,y) \geq 15$, suivant ainsi les paramètres choisis dans [Ferret, 1998]. D'autres liens correspondent en revanche à des relations syntagmatiques[6], telles que « *salaire* » et « *percevoir* », et plus rarement à des relations paradigmatiques[7], comme « *gagner* » et « *perdre* »[8]. Nous établissons le réseau de cooccurrences à partir des années 1999 à 2003 du corpus *Le Monde*, dont on peut espérer qu'il ait une couverture satisfaisante des thèmes abordés dans les émissions que nous étudions, compte tenu de sa nature journalistique et de son grand nombre d'articles.

L'hypothèse d'utilisation de ces types de relations est que les mots qui sont fréquemment voisins ont un lien de sens. Nous prenons en compte ce postulat en augmentant le poids des mots qui présentent des liens avec d'autres mots de la section. Dans l'exemple de la figure 7.6, la pondération affectée à w_2 devient ainsi $f_i(w_2) + coh(w_2, w_5) f_i(w_5) = 1 + 0,2$. Un autre principe qui est souvent employé en TAL pour établir des relations sémantiques entre deux mots est que des termes qui apparaissent dans les mêmes contextes, *i.e.*, au voisinage des mêmes mots, présentent des similarités au niveau du sens. Notre méthode d'acquisition basée sur les cooccurrences ne permet pas d'avoir accès directement à ce type d'informations. Nous proposons toutefois d'étendre les sections supposées par le modèle de segmentation probabiliste par l'ensemble des mots qui lui sont liés d'après le réseau de cooccurrences, en considérant que si ces sections sont homogènes, ce sont les mêmes mots qui sont ajoutés plusieurs fois et qui verront donc leur probabilité d'apparition augmenté dans le modèle de cohésion. En reprenant l'exemple de la figure 7.6, on insère ainsi dans une section S_i le mot w_9, avec une pondération $coh(w_9, w_3) f_i(w_3) + coh(w_9, w_4) f_i(w_4) = 0,21 + 0,18$ dépendant des nombres d'occurences $f_i(w)$ dans S_i des mots auxquels il est lié.

En adoptant une telle démarche, chaque section se trouve ainsi enrichie de mots possédant tous un certain poids, que l'on peut assimiler à des « fractions » d'occurrences. Les nombres de fois où les mots apparaissent deviennent alors des réels, ce qui ne permet

[6] Relations d'ordre sémantique entre deux termes employés au sein de collocations.

[7] Relations d'ordre sémantique entre deux mots présentant une composante commune importante du point de vue du sens, *e.g.* entre deux antonymes.

[8] L'annexe C présente des exemples de relations incluses dans le réseau de cooccurrences.

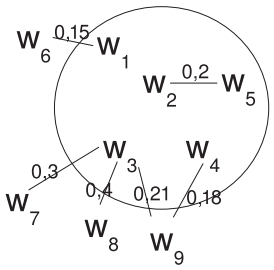

FIG. 7.6: *Extension d'un segment contenant 5 mots en utilisant le réseau de cooccurrences. Les valeurs indiquent la cohésion établie entre deux mots*

taille moy.	rappel_0	précision_0	rappel_2	précision_2	sections homogènes
13,2	44,3	23,3	73,8	38,8	90/118
25,1	47,5	48,3	63,9	65,0	38/62

TAB. 7.4: *Qualité de la segmentation produite sur la transcription automatique des émissions du corpus de test*

plus d'utiliser la loi de Laplace, conçue pour des événements discrets, pour la calcul de $P(W|S)$. Nous avons donc décidé d'employer une loi de probabilité originale, dérivée du modèle unigramme, pour évaluer la probabilité d'apparition de chaque mot $w_j^{(i)}$ présent dans la section S'_i après extension du segment initial S_i :

$$P(w_j^{(i)}) = \frac{f_i(w_j^{(i)}) + \sum_{w \in S_i} f_i(w)\ coh(w, w_j^{(i)})}{n'_i} \tag{7.21}$$

où n'_i est le nombre d'éléments contenu dans S'_i et $f_i(w)$ le nombre d'occurrences de w initialement présent dans S_i. Le principe de calcul de $P(S)$ reste quant à lui le même qu'avant la prise en compte du réseau du cooccurrences.

7.5.2 Résultats et discussion

La comparaison des mesures de la qualité de la segmentation réalisée avec un réseau de cooccurrences (Tab. 7.4) avec celles obtenues avec le modèle initial (Tab. 7.2) montre qu'avec une taille moyenne de sections voisine de 25 groupes de souffle, les valeurs de rappel_0 et précision_0 sont ici plus élevées, ce qui suggère une amélioration apportée par les relations introduites, même si les différences observées demeurent faibles. Le gain constaté est en revanche plus net par rapport à la version utilisant les modèles unigrammes sans faire d'extension (Tab. 7.3).

Les résultats que nous avons présentés démontrent l'intérêt de considérer des relations autres que les simples répétitions de mots puisqu'ils conduisent à une légère amélioration des valeurs de précision et de rappel. La méthode que nous avons proposée n'est toutefois pas entièrement satisfaisante puisqu'elle se limite principalement

aux relations syntagmatiques, autrement dit aux liens qu'entretiennent des mots qui apparaissent souvent ensemble. Même si le réseau de cooccurrences contient quelques relations paradigmatiques, entretenues entre des mots présentant des liens forts au niveau du sens, ce n'est que de manière détournée qu'elles sont prises en compte, en espérant que les extensions de sections conduisent à l'apparition des mêmes mots pour des termes proches au niveau sémantique. Le recours à d'autres méthodes apparaît souhaitable pour savoir si une section est homogène en mesurant le nombre de relations paradigmatiques existant entre ses termes.

La connaissance des liens de proximité sémantique peut être établie à partir de ressources construites manuellement comme des lexiques. Une autre piste très intéressante consiste à employer des méthodes automatiques de construction de classes sémantiques rassemblant des mots de sens proches. Plusieurs techniques ont été développées, notamment pour des corpus spécialisés [Bouaud *et al.*, 1997]. Les méthodes statistiques proposées sur des corpus généralistes constitués d'articles de journaux [Lin et Pantel, 2001 ; Rossignol et Sébillot, 2006] sont applicables dans notre cas pour obtenir des familles de mots similaires au niveau du sens et employés pour traiter un thème particulier. Il serait donc intéressant de tester leur inclusion dans le modèle probabiliste.

Jusqu'à présent, notre méthode ne s'est focalisée que sur l'évolution du nombre d'occurrences de mots pour détecter des ruptures thématiques. Nous souhaitons dans la suite de ce chapitre étudier l'apport d'autres sources d'information, de façon à améliorer la qualité de la segmentation et à identifier plus précisément la localisation des frontières. Un des défauts d'une technique basée sur la seule cohésion lexicale est en effet qu'elle ne permet pas une grande précision sur le repérage des points de rupture thématique, certains groupes de souffle contenant très peu, voire aucun mot lexical.

7.6 Modèle multi-sources

Dans l'objectif de tester la contribution de nouvelles sources d'information pour la segmentation thématique, nous reprenons le modèle initial proposé par Utiyama et Isahara [2001] pour le calcul de $P(W|S)$ et $P(S)$. Nous cherchons dans cette section à étendre ce modèle en considérant deux types de connaissances *a priori* pertinents pour le genre de documents que nous étudions. Le premier regroupe des indices acoustiques, que nous avons restreints à l'alternance de locuteurs homme/femme, à la durée des pauses silencieuses longues et à la durée des *jingles*, qu'ils soient ou non en fond sonore de paroles prononcées. Dans notre objectif de poser des jalons sur la segmentation de ce type de documents, nous nous sommes limité volontairement à des informations simples pouvant être obtenues de manière assez fiable par des procédés automatiques ; ainsi avons-nous préféré l'alternance de locuteurs homme-femme au changement de locuteur, plus délicat à détecter.

Le second type de connaissances correspond au repérage de marqueurs linguistiques, *i.e.*, de suites de mots qui interviennent au voisinage des frontières thématiques. La nature de ces informations est très spécifique au corpus étudié, ne serait-ce qu'à des

chaînes de radio particulières si nous nous restreignons à des émissions d'information. Nous nous intéressons donc également à l'apport des marqueurs constitués de séquences de POS pour avoir une plus grande généricité des connaissances.

Afin de combiner ces différentes sources d'information avec notre modèle de cohésion lexicale, nous proposons un nouveau cadre théorique basé sur une approche probabiliste. Nous déterminons pour ce faire la meilleure hypothèse de segmentation $\hat{S}$ en cherchant celle qui est la plus probable, en ayant observé les trois types de connaissance dont nous disposons :

- W qui est une représentation du document sous la forme d'un « sac de mots », et qui est employée par le modèle de cohésion lexicale,
- F qui forme l'ensemble des caractéristiques acoustiques utilisées,
- M qui est la succession des mots ou de POS du texte analysé et qui permet de repérer des marqueurs linguistiques.

La recherche de $\hat{S}$ s'exprime alors mathématiquement par :

$$\hat{S} = \arg\max_S P(S|W,F,M) = \arg\max_S P(W,F,M|S)P(S) \ . \tag{7.22}$$

Ce calcul de probabilités peut être déterminé à partir de :

$$P(W,F,M|S) = P(W|S)P(F|S)P(M|S) \tag{7.23}$$

si l'on considère que W, F et M sont indépendants. Cette hypothèse nous permet d'employer indépendamment trois modèles prenant en compte chacun une connaissance particulière. Elle est toutefois un peu réductrice. S'il paraît en effet raisonnable de considérer les pauses silencieuses et les successions de mots comme deux modalités différentes, l'hypothèse d'indépendance entre W et M obtenues toutes deux à partir du texte est moins forte. Toutefois, nous considérons ici que W et M ne sont pas liées entre elles dans la mesure où chacune correspond à des informations très différentes. W utilise ainsi une représentation en « sac de mots », en ne conservant que les mots lexicaux, pour mesurer l'homogénéité, tandis que M considère les séquences de POS ou de mots pour repérer des marqueurs linguistiques.

De la même manière que pour notre modèle de cohésion lexicale, nous maintenons l'hypothèse que les frontières thématiques ne peuvent se produire qu'entre deux groupes de souffle. Le cadre proposé pour combiner différentes sources de connaissances revient donc ici à utiliser des MMC multi-flux représentant des MMC indépendants que l'on synchronise à des points fixés, qui sont ici les fins de groupe de souffle [Bourlard et Dupont, 1996].

Nous décrivons dans la suite de cette section les modèles calculant respectivement $P(F|S)$ et $P(M|S)$ et exposons la manière dont nous procédons pour utiliser en pratique l'ensemble des modèles.

7.6.1 Modèle de segmentation à base d'indices acoustiques

Compte tenu de nos hypothèses sur la segmentation, les indices acoustiques auxquels nous nous intéressons ici sont ceux qui interviennent à la fin des groupes de souffle. Les attributs que nous avons retenus sont :

- le fait qu'il se produit ou non une alternance de locuteur homme/femme,
- la durée des pauses silencieuses, en assimilant à 0 toutes celles qui sont inférieures à 1,5 secondes,
- le fait que la fin du groupe de souffle est intervenue ou non durant un *jingle*,
- la durée du *jingle* en cours avant la fin du groupe de souffle,
- la durée du *jingle* en cours après la fin du groupe de souffle.

Ces caractéristiques sont obtenues lors de notre étude des corpus de développement et de test à partir des informations fournies dans l'annotation manuelle du corpus ESTER.

La prise en compte des critères acoustiques intervient dans le calcul de $P(F|S)$ en remarquant que :

$$P(F|S) = \frac{P(S|F)P(F)}{P(S)} \propto \frac{P(S|F)}{P(S)} \tag{7.24}$$

puisque $P(F)$ est une constante lorsque l'on recherche la meilleure segmentation d'un texte.

En supposant que les ruptures thématiques ne se produisent qu'à la fin des groupes de souffle, la recherche de la meilleure segmentation $\hat{S}$ revient à étudier le comportement de $B = B_1 \dots B_\ell$ si ℓ représente le nombre de groupes de souffle contenus dans l'ensemble du document étudié et B_i la variable aléatoire valant « oui » ou « non » selon qu'il existe ou non une rupture thématique entre le i-ème groupe de souffle et le $(i+1)$-ème. Pour B_ℓ, il s'agit d'une rupture à la fin du texte. Le calcul de $P(S|F)$ peut alors se faire de la manière suivante :

$$P(S|F) = P(B|F) = \prod_{i=1}^{\ell} P(B_i|F) \ . \tag{7.25}$$

si l'on considère que la décision prise pour chaque frontière B_i est indépendante des autes. D'où, en notant les caractéristiques acoustiques se produisant au voisinage de B_i par F_i et en faisant l'hypothèse que B_i ne dépende que des caractéristiques acoustiques F_i dans son environnement immédiat par :

$$P(S|F) = \prod_{i=1}^{\ell} P(B_i|F_i) \ . \tag{7.26}$$

Nous avons décidé d'estimer $P(B_i|F_i)$ au moyen d'un arbre de décision. Cette méthode de classification, couramment utilisée y compris pour la segmentation thématique [Tür *et al.*, 2001], a été retenue ici pour deux raisons principales. Elle nous offre tout d'abord une visualisation simple du modèle de détection de changement de thèmes qui nous permet de l'interpréter. En outre, l'utilisation de l'algorithme C4-5 [Quinlan, 1993] nous permet d'envisager des attributs aussi bien numériques que symboliques comme critères.

Nous établissons l'arbre de décision, grâce au logiciel libre WEKA [Witten et Frank, 2005], sur l'ensemble d'apprentissage du corpus ESTER, dont les émissions sont découpées en entités journalistiques. Cette segmentation diffère de celle que nous avons faite nous-même sur le corpus de développement mais elle nous permet d'avoir accès à un nombre beaucoup plus important de frontières entre segments. Les probabilités

$P(B_i =$ « oui ») et $P(B_i =$ « non ») sont déterminées en comptant la proportion d'éléments des classes « oui » et « non » au niveau de chaque feuille de l'arbre.

Nous obtenons un arbre contenant 13 nœuds, dont 7 feuilles. De manière assez surprenante, l'information sur la durée des pauses silencieuses n'a pas été prise en compte pour prédire la présence de frontière thématique. En examinant de plus près notre corpus d'étude, nous avons pu constater que les longs moments de silence n'étaient pas le fait des journalistes, mais plutôt de personnes qu'ils interrogeaient et sujettes à des hésitations, ce qui explique pourquoi leur présence soit aussi peu corrélée avec le passage d'une entité journalistique à une autre. C'est l'alternance de locuteurs homme/femme qui a été jugée comme critère le plus discriminant, suivie par la présence de musique. Les connaissances d'ordre acoustique constituent des éléments intéressants pour positionner des points de rupture thématiques. Néanmoins, ils ne sont pas toujours très précis, des *jingles* pouvant se produire au niveau de plusieurs groupes de souffle.

7.6.2 Modèle de segmentation à base de marqueurs linguistiques

Nous souhaitons ici détecter des marqueurs de discours, indicateurs de frontières thématiques, à partir d'une transcription. Ces indices peuvent se présenter sous la forme de séquences de mots telles que « *France-info sports* », voire une suite de POS après étiquetage, en supposant qu'après certains modes de formulation, il existe une plus forte probabilité de voire apparaître une nouvelle section. Afin de déterminer automatiquement les marqueurs utiles dans notre contexte et de les employer au sein du cadre théorique que nous avons défini, nous recourons aux modèles N-grammes cachés. Nous souhaitons montrer l'intérêt de ces outils — utilisés habituellement en vue de repartitionner en phrases une transcription découpée en groupes de souffle [Stolcke *et al.*, 1998] — pour segmenter thématiquement un document.

De manière similaire aux indices acoustiques, nous obtenons $P(M|S)$ par :

$$P(M|S) \propto \frac{P(S|M)}{P(S)} \tag{7.27}$$

avec :

$$P(S|M) = P(B|M) = \prod_{i=1}^{\ell} P(B_i|M) \ . \tag{7.28}$$

Le recours aux modèles N-grammes cachés intervient alors au niveau du calcul de $P(B_i|M)$. Le principe de cette technique consiste à faire correspondre à chaque mot w du texte une variable correspondant à un événement dit caché et à employer le ML N-grammes caché pour prédire cet état pour chaque mot d'un groupe de souffle. Nous adaptons cette méthode à notre contexte applicatif en considérant que la variable cachée peut être associée à deux états : « oui » pour marquer une rupture thématique et « non » dans le cas contraire. La construction du modèle pour la prise en compte des marqueurs linguistiques sous forme de mots est réalisée à partir de l'ensemble des émissions extraites du corpus d'apprentissage d'Ester et représentées chacune sous

la forme « $<s>w_1 b_1 w_2 b_2 \ldots w_i b_i <eob> b_{i+1} w_{i+2} b_{i+2} \ldots </s>$ » où les b_i prennent la valeur « *oui* » ou « *non* » selon que le mot w_i est suivi ou non par une rupture thématique et où $<eob>$ est un marqueur spécial, assimilé à un mot et indiquant une fin de groupe de souffle. L'apprentissage des ML de ce type permet d'estimer les probabilités $P(b_i|w_1 b_1 \ldots w_{i-1} b_{i-1} w_i)$ et $P(w_i|w_1 b_1 \ldots w_{i-1} b_{i-1})$. Celles-ci sont utilisées pour prédire les événements cachés, autrement dit absents du texte original, à partir des seules séquences de mots $<s>w_1 \ldots w_i <eob> w_{i+2} \ldots w_n </s>$. Ne cherchant pour notre part qu'à déterminer $P(B_i|M)$ à la fin de chaque groupe de souffle, nous restreignons nos calculs aux probabilités *a posteriori* d'observer une rupture thématique après les marqueurs $<eob>$, en obtenant ces valeurs à l'aide d'un algorithme *forward-backward*.

La construction et l'utilisation du ML établi pour les POS sont basées sur le même principe. La seule différence réside dans le fait que nous étiquetons au préalable chaque émission et considérons des étiquettes t_i au lieu des mots w_i.

Nous établissons deux ML 6-grammes cachés à l'aide de la boîte à outils SRILM [Stolcke, 2002], l'un pour les mots, l'autre pour les POS[9], ce qui revient à considérer un historique $w_{i-2} b_{i-2} w_{i-1} b_{i-1} w_i$ constitué de 3 mots (ou POS) précédents et de deux éléments cachés. Le ML basé sur les mots permet notamment de prédire une rupture thématique avec une probabilité élevée quand le groupe de souffle se termine par « *excellente journée* », « *semaine prochaine* », *huit heures* » ou « *point com* », ou quand il débute par « *c'était* » ou « *bonjour à* ». Celui basé sur les POS conduit à des probabilités plus élevées après des groupes de souffle finissant par des noms propres ou par un cardinal suivi par un nom commun féminin pluriel — correspondant à des formules du style « *neuf heures* ».

Nous avons vu au cours des sections précédentes le principe de chacun des modèles associé à une modalité que nous prenons en compte dans la segmentation. Nous présentons dans ce qui suit comment ces techniques sont agencées de manière à ce que le découpage d'un texte se fasse en un temps raisonnable.

7.6.3 Méthode d'implémentation

D'après la description des différents modèles que nous avons décrits et notre approche générale, l'objectif de la segmentation s'exprime sous la forme

$$\hat{S} = \arg\max_S P(W|S) P(F|S) P(M|S) P(S) \tag{7.29}$$

$$= \arg\max_S P(W|S) \frac{P(S|F)}{P(S)} \frac{P(S|M)}{P(S)} P(S) \ . \tag{7.30}$$

Nous introduisons trois coefficients de pondération α, β et γ qui permettent de tenir compte des différences de facteur d'échelle entre les différentes probabilités. Les valeurs calculées pour $P(B_i = \text{« oui »}|F_i)$ et $P(B_i = \text{« oui »}|M)$ sont en particulier très faibles puisque la très grande majorité des fins de groupe de souffle ne sont pas associées à des frontières thématiques. En outre, ces coefficients permettent de moduler la pertinence

[9] Nous avons recouru au jeu d'étiquettes *ext2*, *i.e.*, à celui que nous avons employé pour réordonner les listes d'hypothèses au chapitre 5.

des modèles vis-à-vis de la tâche de segmentation. En conséquence, de manière finalement similaire au problème de la RAP (*cf.* équation (1.11) page 16), nous redéfinissons l'objectif de la segmentation par :

$$\hat{S} = \arg\max_{S} P(W|S)P(S|F)^{\alpha}P(S|M)^{\beta}P(S)^{\gamma} \ . \tag{7.31}$$

Du point de vue de l'implémentation de ce nouveau modèle, celle-ci peut s'effectuer de manière similaire au modèle de cohésion lexicale en cherchant :

$$\hat{S} = \arg\min_{S} -\log\left[P(W|S)P(S|F)^{\alpha}P(S|M)^{\beta}P(S)^{\gamma}\right] \ . \tag{7.32}$$

La valeur d'un coût d'un segment S_i constitué des groupes de souffle $s_a \dots s_b$ devient :

$$\begin{aligned} C(S_i|W,F,M) \quad = \quad & -\log P(W_i|S_i) \\ & -\alpha\left[\sum_{j=a}^{b-1}\log P(B_j = \text{« non »}|F_j) + \log P(B_b = \text{« oui »}|F_b)\right] \\ & -\beta\left[\sum_{j=a}^{b-1}\log P(B_j = \text{« non »}|M) + \log P(B_b = \text{« oui »}|M)\right] \\ & -\gamma\log P(S) \ . \end{aligned} \tag{7.33}$$

Ce calcul de coût permet alors de construire un graphe de même topologie que celui employé pour le modèle probabiliste initial et donc de recourir à l'algorithme de programmation dynamique similaire à celui déjà présenté.

7.6.4 Évaluation expérimentale

Nous mesurons l'influence des différentes sources d'information en les ajoutant progressivement au modèle de cohésion lexicale. Nous rappelons dans la première partie du tableau 7.5 les résultats que nous avons obtenus sur le corpus de test avec la méthode probabiliste initiale. Nous recourons au corpus de développement pour fixer les valeurs des coefficients de pondération α et β, en choisissant comme critère d'optimisation la maximisation du rappel_0 pour une taille moyenne de sections voisine de celle de la segmentation de référence, égale à $23,3$ groupes de souffle. Cette décision est un peu arbitraire, d'autres indices comme la F-mesure — prenant en compte à la fois la précision et le rappel — ou rappel_1 pouvant être tout aussi viables. γ est lui aussi réglé sur le corpus de développement mais en retenant cette fois-ci deux valeurs, conduisant respectivement à des segments de longueurs moyennes d'environ 25 et 15 groupes de souffle. Diminuer la taille des sections nous permet en effet d'augmenter le rappel, même si cela se fait au détriment de la précision. Nous avons de plus constaté que les nouveaux points de rupture permettent assez souvent d'isoler les *fillers* des autres sections et ne morcellent donc pas outre mesure les segments traitant un thème particulier.

Le tableau 7.5 présente les valeurs de précision et de rappel calculées en utilisant en sus de la cohésion lexicale, les marqueurs linguistiques de mots (deuxième partie), les

taille moy.	rappel_0	précision_0	rappel_2	précision_2	sections homogènes
cohésion lexicale					
15,7	57,4	36,1	80,3	67,0	78/99
22,2	44,3	39,7	72,1	66,2	46/70
cohésion lexicale + marqueurs de POS					
16,6	62,3	41,3	83,6	65,2	75/94
25,5	50,8	52,5	70,5	72,9	40/61
cohésion lexicale + marqueurs de mots					
17,7	57,4	38,1	85,3	67,4	72/94
25,5	50,8	52,5	70,5	76,3	36/61
cohésion lexicale + indices acoustiques					
17,7	62,3	44,2	77,1	61,6	68/88
25,5	59,0	61,0	72,1	81,4	40/61
cohésion lexicale + marqueurs de POS + indices acoustiques					
16,1	67,2	43,2	85,3	65,3	81/97
23,6	59,0	56,3	75,4	79,7	40/61
cohésion lexicale + marqueurs de POS + marqueurs de mots + indices acoustiques					
16,2	65,6	42,6	88,5	64,9	79/96
25,1	60,7	61,7	78,7	81,7	42/62

TAB. 7.5: *Qualité de la segmentation produite sur la transcription automatique des émissions du corpus de test en utilisant diverses sources d'information*

marqueurs linguistiques de POS (troisième partie), les indices acoustiques (quatrième partie), les indices acoustiques et les marqueurs de POS (cinquième partie) et enfin l'ensemble des indices disponibles (dernière partie). Le lecteur pourra se faire une idée plus précise de la qualité de la méthode automatique en examinant dans l'annexe D un extrait d'une émission de corpus de test découpé selon la technique employée dans la quatrième partie du tableau 7.5 — avec une taille moyenne de 16,1 groupes de souffle).

Afin d'utiliser simultanément les marqueurs à base de POS M_t et ceux constitués de mots M_w, nous avons fait l'approximation que ces deux informations sont indépendantes :

$$P(B_i|M) = P(B_i|M_t, M_w) \approx \frac{P(B_i|M_t)P(B_i|M_w)}{P(B_i)} \tag{7.34}$$

Bien entendu, cette hypothèse n'est guère valide puisque les deux catégories de marqueurs sont très liées ; elle nous permet toutefois de ne pas introduire un nouveau coefficient de pondération des deux probabilités $P(B_i|M_t)$ et $P(B_i|M_w)$, tout en obtenant des résultats voisins de ceux calculés avec une interpolation entre les deux modèles, selon des essais que nous avons effectués.

La comparaison de la qualité de la segmentation obtenue dans différentes configurations montre que toutes les nouvelles connaissances introduites permettent d'améliorer

de manière prononcée les valeurs de précision et de rappel, ce qui confirme l'intérêt de notre approche pour combiner différentes sources d'information. Le nombre de segments homogènes est lui aussi augmenté mais de façon beaucoup moins nette. Un examen plus approfondi conduit à remarquer que de manière assez surprenante les marqueurs de POS conduisent à des résultats similaires à ceux obtenus avec des marqueurs de mots. Ceci montre que des suites de POS sont des indicateurs tout à fait pertinents pour identifier des changements de thème. En confrontant l'apport des marqueurs linguistiques avec celui des indices acoustiques, il apparaît que ces dernières conduisent à un accroissement plus net de la précision et du rappel, suggérant ainsi que ce second type de connaissance est plus informatif. La combinaison de l'ensemble des modèles dont nous disposons correspond globalement aux meilleurs résultats que nous obtenons, ce qui souligne leur complémentarité.

7.7 Bilan et discussion

D'après les résultats que nous avons présentés, la segmentation thématique de transcriptions automatiques reste une tâche difficile, d'autant plus qu'il est délicat d'évaluer précisément la qualité des découpages obtenus en fonction de notre cadre applicatif. En conclusion de ce chapitre prospectif, nous réalisons un bilan de nos différentes techniques de segmentation avant de discuter de leurs améliorations possibles et des évolutions nécessaires pour une intégration pleinement opérationnelle dans la chaîne de traitements que nous avons décrite au chapitre 6.

7.7.1 Qualité des résultats obtenus

Même si les valeurs de précision et de rappel fournies restent assez éloignées des 100 % souhaités, les mesures effectuées de la qualité de la segmentation nous indiquent des résultats relativement satisfaisants. En choisissant ainsi d'avoir un découpage conduisant à des sections un peu plus petites que celles des émissions segmentées manuellement et en tolérant des erreurs à deux groupes de souffle près sur les frontières trouvées, la méthode de cohésion lexicale employée permet ainsi de détecter plus de quatre ruptures thématiques sur cinq, tout en proposant des frontières qui se révèlent dans deux cas sur trois corrects (Tab. 7.2 page 134). Ces évaluations sont encourageantes, d'autant plus que la métrique que nous utilisons considère comme incorrectes des frontières pourtant bien reconnues entre sous-thèmes. Plus important encore, parmi les sections proposées par notre méthode automatique, 80 % des sections ne traitent que d'un seul thème. Même si certains de ces segments sont courts, en ne comportant que deux à trois groupes de souffle, le fait d'obtenir une large majorité d'unités homogènes est rassurant pour les étapes suivantes de l'adaptation du ML. En examinant manuellement le découpage proposé, nous avons en outre remarqué que les tailles des sections étaient très diverses, tout comme l'étaient celles de la segmentation manuelle de référence, ce qui montre ainsi la capacité du modèle à s'adapter au rythme avec lequel les sujets sont traités. Toutes ces constatations soulignent l'intérêt de la méthode choisie

pour nos types de documents, pour peu que l'on ait une idée sur la taille moyenne souhaitée des sections. Le découpage en groupes de souffle, qui sont des unités de longueur généralement beaucoup plus réduites que les phrases, ne semble pas perturber le modèle probabiliste. De même, les erreurs de reconnaissance, bien qu'elles soient en nombre non négligeable comme l'attestent nos calculs du LER, ne semblent pas rédhibitoires.

Des expériences complémentaires menées sur l'intégration d'autres connaissances que la cohésion lexicale montrent une amélioration de la détection des points de rupture. En étendant ainsi les mots lexicaux de la transcription par des mots auxquels ils sont sémantiquement liés, nous avons obtenu une légère augmentation des valeurs de précision et de rappel. Ces résultats ont été produits à partir de la méthode par défaut proposée par Ferret [1998] pour sélectionner les cooccurrences à retenir. Il est probable que l'on puisse obtenir de meilleurs résultats en utilisant des connaissances plus poussées sur les relations sémantiques paradigmatiques.

L'emploi des caractéristiques acoustiques permet lui aussi de guider les choix de rupture thématique. L'arbre de décision construit pour ce faire remet en outre en question, au moins pour le type de documents que nous avons étudié, la pertinence des pauses silencieuses pour détecter le passage d'un sujet à un autre. Dans les émissions que nous avons analysées, les moments prolongés de silence sont parfois utilisés par le journaliste pour marquer une pause avant d'évoquer ensuite le sujet suivant. Toutefois, l'apparition des nouveaux sujets est plus fréquemment marquée par l'intervention d'un *jingle* ou par la prise de parole de la part d'un autre locuteur, ce qui confère un rythme plus dynamique à l'émission. Les pauses silencieuses se trouvent ainsi au final plutôt liées à la présence de disfluences.

La prise en compte des marqueurs linguistiques au moyen de N-grammes cachés établit également l'intérêt de cette méthode au vu de l'amélioration constatée, aussi bien en repérant les séquences de POS que les suites de mots. L'apport des informations morpho-syntaxiques est particulièrement encourageant car cette connaissance semble beaucoup moins spécifique que le repérage de séquences de mots particulières.

La combinaison de tous ces indices conduit à une amélioration significative des résultats puisqu'en tolérant des erreurs à deux groupes de souffle près, la précision et la valeur de rappel atteignent simultanément des valeurs voisines de 80%. Cette constatation souligne d'une part tout l'avantage de combiner différentes sources d'information et valide d'autre part la méthode proposée pour intégrer ces différentes sources.

Malgré tous ces points positifs, la segmentation obtenue est encore défectueuse à certains endroits de la transcription. Les sections les plus problématiques sont incontestablement celles présentant les titres d'une émission, responsables de la majorité des erreurs de détection de ruptures thématiques. Dans les documents que nous avons étudiés, les points principaux sont répétés jusqu'à cinq fois au cours d'une heure de retransmission. Ces événements récurrents posent deux difficultés pour notre méthode basée sur la cohésion lexicale. Ils sont, d'une part, constitués d'une suite de thèmes qui sont chacun abordés en un à trois groupes de souffle, introduisant ainsi une variation locale inhabituelle des probabilités mesurant l'homogénéité des sections. D'autre part, ils abordent par définition des thèmes qui sont traités par la suite. Ceci conduit parfois à regrouper les titres avec le premier, voire les deux premiers sujets leur succédant, la

technique de segmentation employée détectant en effet des mots déjà apparus dans les titres dans la suite du texte.

7.7.2 Améliorations possibles de la segmentation thématique

Les titres paraissent difficiles à traiter spécifiquement, à moins d'introduire des règles *ad hoc* sur le fait qu'ils interviennent avec un fond musical et sur les instants où ils se produisent au cours d'une émission d'information, ceux-ci étant espacé de manière régulière. Cette méthode présente l'inconvénient évident d'être très spécifique à chaque type de programme étudié et c'est donc plutôt une amélioration globale de la méthode de segmentation qu'il faut viser.

Une des manières de faire progresser les résultats consiste à poursuivre l'intégration de connaissances que nous avons initiée. Comme nous l'avons déjà mentionné à la section 7.5.2, la prise en compte des relations paradigmatiques nous paraît intéressante pour tenir compte de la richesse lexicale pour traiter d'un sujet. En particulier, le recours à des classes sémantiques élaborées à partir de sous-corpus thématiques et donc possédant une coloration thématique [Rossignol et Sébillot, 2006] nous paraît adapté à notre tâche puisque le repérage de plusieurs mots de telles classes dans une même section semble un indicateur fort d'homogénéité de ce segment.

Nous avons déjà souligné l'intérêt de l'alternance des locuteurs homme/femme (*cf.* section 7.6.1) pour détecter les points de rupture thématique. L'étude du changement de locuteur au moyen des méthodes automatiques doit *a priori* apporter un gain au niveau de la segmentation thématique.

Une autre amélioration souhaitable concerne la prise en compte des erreurs de transcription. Si nous avons en effet remarqué que certaines erreurs sont réalisées de manière répétée, *e.g.* « *bernard tapir* » à la place de « *bernard tapie* », ce qui ne pénalise pas notre méthode basée sur les répétitions de mots, la mauvaise reconnaissance de certains mots porteurs de sens est plus problématique. L'utilisation des indices de confiance devrait permettre de prendre en compte ce phénomène.

Enfin, un dernier type d'erreurs de segmentation a retenu notre attention et paraît assez facilement corrigeable. Nous avons constaté qu'en recourant à des modèles établis à partir d'un corpus découpé en entités journalistiques, certaines parties du document sont sursegmentées. L'exemple typique correspond au cas où un reportage annoncé par un journaliste est considéré comme une section à part entière alors qu'il n'intervenait que pour illustrer un sujet traité. Ce genre d'erreurs semble rectifiable en examinant après une premier découpage la similarité lexicale entre deux sections voisines.

7.7.3 La segmentation en vue d'adapter un ML

Certains segments proposés sont encore trop approximatifs pour pouvoir être utilisés tels quels par les étapes suivantes de l'adaptation du ML. Il semble donc nécessaire de développer de nouvelles méthodes permettant d'écarter les sections qui sont non pertinentes pour constituer ensuite des corpus d'adaptation de qualité.

Deux critères de sélection nous semblent importants pour juger de la qualité d'un segment pour la tâche que nous nous sommes fixée. Le premier concerne la taille des sections. Il est en effet nécessaire de s'assurer que le segment proposé soit suffisamment long pour pouvoir extraire suffisamment de mots-clés pour formuler des requêtes en vue de la formation de corpus. Ceci peut être contrôlé de manière simple en limitant l'adaptation à des sections qui contiennent un nombre de mots lexicaux supérieur à un seuil fixé, par exemple une vingtaine.

Le second critère doit mesurer la cohésion d'une section thématique, de manière à ne retenir que celles qui traitent d'un seul sujet. Ceci doit notamment permettre d'écarter les titres d'émission, si problématiques pour notre approche d'adaptation thématique. Cet indice souhaité peut être vu comme une mesure de désordre ou de dispersion lexicale. Les méthodes dérivées de l'entropie, définie en théorie de l'information, nous semblent pouvoir constituer un critère possible d'élimination. Dans notre cas, cette mesure pourrait être mesurée pour une section donnée S_i sous la forme :

$$H = -\sum_{w \in S_i} P(w) \log P(w) \tag{7.35}$$

où $P(w)$ peut être établie de manière similaire au calcul de l'homogénéité d'une section, utilisé dans le modèle de cohésion lexicale. Une autre notion intéressante est celle de l'énergie textuelle [Fernández *et al.*, 2007]. Les sections traitant de différents sujets pourraient alors être assimilées à des zones pour lesquelles le spectre mesurant l'énergie d'une phrase est activée pour plusieurs valeurs éloignées.

Au terme de ce chapitre, nous avons ainsi étudié l'influence de différentes sources d'information sur la qualité de la segmentation thématique d'une transcription. Si les valeurs de précision et de rappel que nous mesurons à partir de découpages produits automatiquement sont assez satisfaisantes, certaines sections paraissent encore trop hétérogènes pour que notre méthode de segmentation soit utilisée telle quelle dans la chaîne de traitements que nous avons proposée pour adapter le ML. Aussi considérons-nous des sections sélectionnées manuellement et donc supposées homogènes pour nos travaux décrits dans le chapitre suivant. Celui-ci est consacré aux dernières étapes de l'adaptation, à savoir la formation de corpus issus du Web puis la construction de ML adaptés pour chaque section thématique.

Chapitre 8

Adaptation du modèle de langage par des corpus issus du Web

Le but d'un modèle de langage en RAP est d'assigner des probabilités à des séquences de mots afin d'aider à la désambiguïsation d'hypothèses similaires sur le plan acoustique. Or, les systèmes de transcription ne disposent pas toujours d'un ML qui corresponde exactement au type de document étudié. Celui-ci peut concerner un domaine spécialisé inconnu ou tout du moins mal connu par le ML ; le ML peut également être construit de manière à avoir une large couverture d'une langue, au risque d'être trop général et de moins bien modéliser un thème particulier que des ML spécifiques. Le mécanisme d'adaptation vise à résoudre ces difficultés en modifiant le ML pour correspondre au mieux aux données à transcrire.

Dans ce chapitre, nous supposons que chaque section thématique étudiée, que nous assimilons dorénavant à un document à part entière, est suffisamment uniforme pour être traité comme un tout. Même en faisant ce postulat, l'adaptation soulève deux difficultés principales. Une première est de disposer d'assez de données pertinentes pour construire un ML pour chaque domaine étudié. Contrairement à de nombreuses approches [Klakow, 2000 ; Chen *et al.*, 2001], nous ne nous limitons pas pour cette raison à un corpus pour obtenir les informations nécessaires à la constitution des ML ; nous avons plutôt choisi de recourir au Web, dont l'utilisation est moins répandue. Le principal intérêt de cette gigantesque source d'information est qu'elle permet de couvrir de manière assez exhaustive les thèmes qui peuvent être rencontrés dans le document étudié, surtout pour des sujets relevant de l'actualité récente. Beaucoup d'études faisant appel au Web exploitent cette ressource pour obtenir des corpus sur un domaine comme celui de la médecine [Sethy *et al.*, 2005 ; Suzuki *et al.*, 2006] ou sur un registre comme celui des conversations spontanées [Bulyko *et al.*, 2007]. Nous l'utilisons quant à nous pour construire des corpus pertinents du point de vue des sujets abordés pour chaque document à transcrire.

Le second point délicat à traiter concerne la sélection des données constituant les corpus d'adaptation. La plupart des études qui ont été menées jusqu'à présent sur ce sujet s'appuient sur une liste fixée *a priori* de thèmes, ce qui ramène ce problème à savoir détecter le ou les sujets qui sont traités dans le document courant [Gildea et Hofmann, 1999 ; Iyer et Ostendorf, 1999]. Une alternative également considérée repose sur l'exploitation d'un corpus — généralement de taille modeste — représentatif du domaine étudié pour formuler des requêtes à un moteur de recherche sur Internet [Bulyko *et al.*, 2003 ; Sethy *et al.*, 2005]. Nous avons décidé quant à nous de ne pas recourir à de telles listes fermée de thèmes, ni d'employer de textes spécialisés, ce qui nous permet d'adapter la RAP pour n'importe quel sujet pouvant être rencontré dans un document. Nous proposons pour ce faire une approche originale qui repose sur la détection de mots jugés suffisamment représentatifs du document traité, à partir d'une première transcription produite par le système de RAP. Ces mots-clés sont employés pour construire des requêtes qui sont ensuite formulées à un moteur de recherche sur le Web (Fig. 8.1). Cette démarche nous a conduit à concevoir de nouvelles méthodes pour s'assurer de la qualité des corpus construits. Ainsi, bien que des techniques existent pour détecter les mots les plus caractéristiques d'un document, notamment dans le domaine de la RI, elles sont rarement employées sur des textes aussi bruités que des transcriptions automatiques. Nous proposons donc de prendre en compte la mesure de confiance dans le calcul du critère de sélection des mots-clés pour traiter le cas des mots mal reconnus. Nous contrôlons en outre *a posteriori* que les pages Web fournies en réponses par le moteur de recherche correspondent bien au texte traité.

Dans la lignée des deux chapitres précédents qui se veulent avant tout exploratoires, nous souhaitons montrer ici avant tout l'intérêt du Web pour construire des ML spécifiques à un document, et l'adéquation des méthodes de RI pour réaliser cette tâche. Si cette source d'information est indéniablement très riche, elle est aussi très hétérogène. Nous nous sommes donc principalement consacré à élaborer un prototype permettant de constituer des corpus de qualité satisfaisante et surtout pertinents par rapport aux documents traités. Nous n'avons pas cherché en revanche à utiliser des méthodes de conception plus complexes pour adapter le ML, la simple utilisation d'une interpolation linéaire nous permettant déjà de mesurer l'apport de notre démarche pour améliorer la modélisation du langage.

La première section de ce chapitre est consacrée aux méthodes que nous proposons pour obtenir un ensemble de données en provenance du Web et spécifique à chaque document étudié[1]. Nous montrons ensuite que des ML construits à partir de corpus d'adaptation ainsi constitués améliorent la qualité de la transcription de manière modérée. Nous présentons enfin dans une section bilan les perspectives ouvertes par nos travaux.

[1] Les étapes de détection de mots-clés et de constitution de corpus à partir de pages Web ont été réalisées par Gwénolé Lecorvé lors de son stage de master recherche [Lecorvé, 2007].

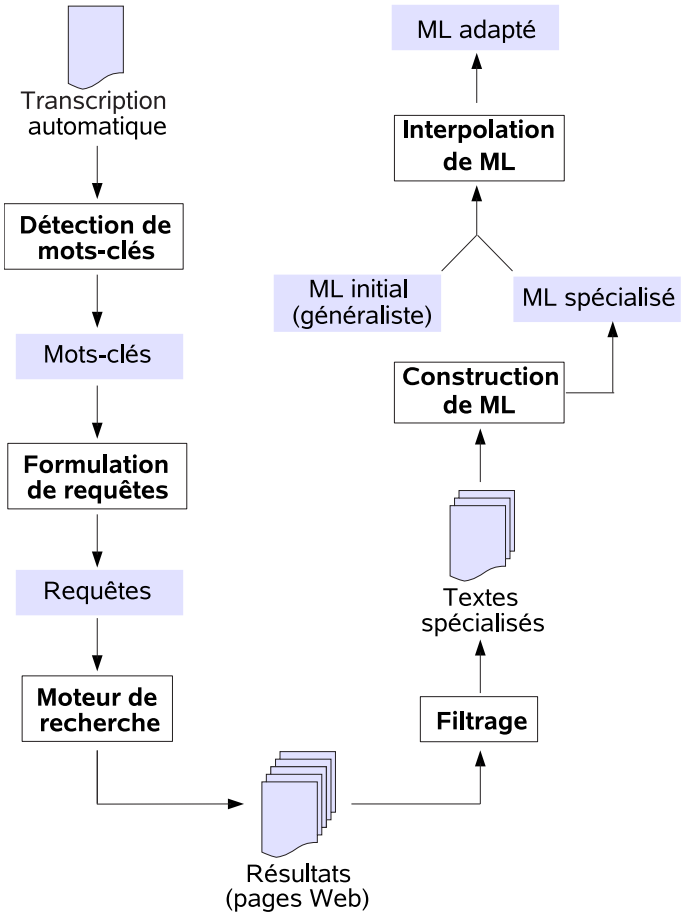

FIG. 8.1: *Construction de corpus d'adaptation*

8.1 Construction de corpus à partir d'Internet

Le Web est ici considéré comme une très riche source d'information que nous souhaitons exploiter pour constituer des corpus traitant principalement de ce dont « parle » un document à transcrire. Pouvoir repérer les textes intéressants parmi l'ensemble des pages Web disponibles sur Internet pose plusieurs défis. Le premier d'entre eux consiste à savoir de quoi traite une transcription. Pour le résoudre, nous extrayons un ensemble de mots-clés jugés représentatifs selon un critère statistique que nous avons choisi, et ce pour chaque document étudié. Le deuxième problème qui apparaît concerne la récupération de données pertinentes. Nous employons pour ce faire un moteur de recherche grand public sur Internet en formant des requêtes au moyen des mots-clés précédemment repérés dans la transcription, puis en consultant les pages Web fournies parmi les premières pages de résultats. Le dernier point à fixer est la constitution de corpus à partir des réponses produites par le système de RI. Les données trouvées peuvent

notamment être de qualité variable et il convient de s'assurer de leur adéquation avec le document traité.

La suite de cette section décrit notre méthode pour concilier l'ensemble de ces difficultés, en relatant successivement le critère choisi pour sélectionner les mots-clés, le mode de formulation des requêtes au moteur de recherche et enfin la sélection des données pertinentes parmi les pages retournées en résultats.

8.1.1 Extraction de mots-clés

La constitution de corpus d'adaptation nécessite, dans notre système, l'extraction automatique des mots-clés pour chaque document. Cette étape est particulière à notre approche puisque beaucoup d'autres méthodes d'adaptation s'appuient sur une liste de thèmes préétablie pour décider du ou des ML spécifiques à un sujet particulier à appliquer. Le principal avantage de notre méthode est qu'en utilisant directement les informations présentes dans le texte à traiter, nous ne faisons pas d'hypothèses sur les domaines pouvant apparaître dans les transcriptions rencontrées.

Afin de déterminer quels mots sont mieux à même de représenter le contenu thématique d'un document, nous faisons l'hypothèse simplificatrice que ce sont ceux qui apparaissent souvent dans la transcription et peu dans d'autres textes qui sont les plus caractéristiques. Cette méthode ne permet pas de s'assurer que ces termes traitent véritablement du sujet principal ; en restreignant le choix des mots-clés à 5, nous pouvons toutefois espérer qu'ils soient suffisamment importants vis-à-vis du contenu sémantique du document pour que d'autres pages Web qui les contiennent abordent des thèmes similaires. Afin de mesurer l'importance d'un terme dans la transcription étudiée, nous recourons au critère du *tf-idf* (pour *term frequency* et *inverse document frequency*) qui a largement fait ses preuve en RI [Salton, 1989]. Mathématiquement, la valeur $\sigma(w)$ que nous calculons pour un terme w présent dans le texte T et en disposant d'un corpus C constitué de nombreux documents s'obtient par :

$$\sigma(w) = tf(w,T)\, idf(w) = \frac{|w|_T}{|T|} \log\left(\frac{|C|}{Card\,\{d \in C | w \in d\}}\right) \tag{8.1}$$

où $|w|_T$ est le nombre d'occurrences de w dans T. Durant nos expériences, nous avons assimilé le corpus d'apprentissage d'Ester au corpus de référence C, en considérant que chaque segment associé à une entité journalistique formait un document.

Certains termes, essentiellement les mots grammaticaux, peuvent être éliminés avec certitude des prétendants pour figurer dans la liste des mots-clés. De manière similaire à notre modèle de segmentation basé sur la cohésion lexicale, nous ne conservons donc que les noms communs, les noms propres, les adjectifs et les verbes (*cf.* section 7.3). Certains mots morphologiquement très proches peuvent en outre être considérés comme semblables au niveau du sens. Nous recourons à un *stemmer* pour le français[2] pour ôter les suffixes de chaque terme. Les *stems* ainsi obtenus sont parfois des troncatures de mots. Nous remplaçons donc lors de la construction de requêtes chaque *stem* s associé

[2]Disponible sous la forme d'un module *Perl*, nommé `Lingua::Stem::Fr`.

à un haut score *tf-idf* $\sigma(s)$ par le terme qui présente le plus d'occurrences dans le texte original, parmi l'ensemble des mots qui lui sont associés (Fig. 8.2). Nous faisons ainsi l'hypothèse que ce mot est représentatif du *stem* dont il est dérivé et que les autres qui lui correspondent apportent des informations redondantes. Cette dernière opération est rendue nécessaire par le fait que le moteur de recherche employé, la version française de *Yahoo !*, n'est pas conçu pour recevoir des racines parmi ses requêtes.

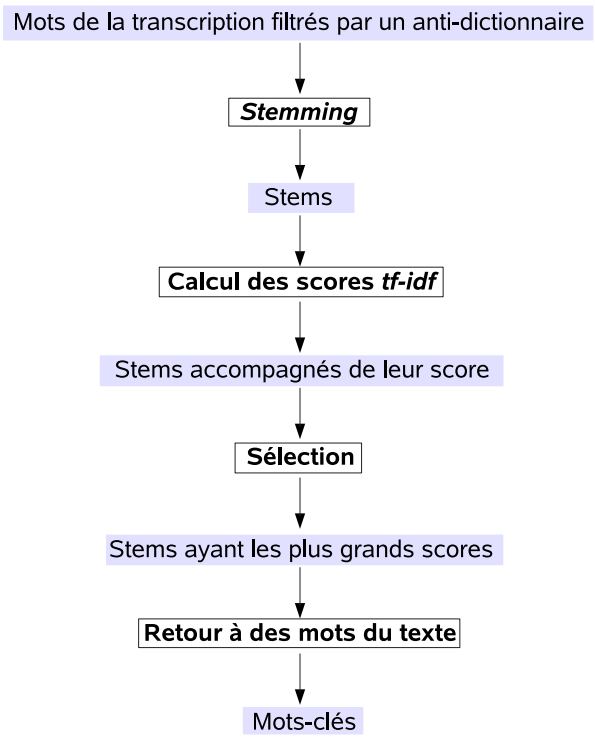

FIG. 8.2: *Sélection des mots-clés*

L'examen manuel des mots-clés sélectionnés grâce au critère classique *tf-idf* sur quelques documents montre qu'ils résument souvent assez bien les textes dont ils sont extraits (*cf.* annexe E). Le repérage de mots relatifs au champ lexical de la maladie, tels que « *cancer* », « *psychopatologie* » ou « *malade* », pour un reportage traitant du « plan cancer » permet par exemple de trouver des documents qui sont liés à ce sujet. Nous avons en revanche trouvé des cas plus problématiques où les mots-clés choisis ne laissaient pas espérer que les corpus construits par la suite fussent pertinents. Les termes sélectionnés peuvent alors être trop spécifiques, en concernant par exemple des noms de lieux ou de personnes, mentionnés dans la transcription, mais anecdotiques par rapport au contenu sémantique du document. Ils peuvent également être complètement hors sujet, en particulier du fait de mots mal reconnus. Pour limiter ces cas de figures, nous proposons deux modifications du score *tf-idf* $\sigma(s)$.

Traitement des noms propres

Les noms propres sont problématiques sur deux plans. D'une part, ils ne permettent pas toujours de représenter convenablement un document, alors que leur emploi rare dans le corpus de référence leur confère une haute valeur pour l'*idf*. Ils conduisent d'autre part à réduire le nombre de réponses fournies par le moteur de recherche lorsque l'on en combine plusieurs — ne serait-ce que deux ou trois — pour constituer des requêtes. Nous introduisons donc une pénalité sur le score *tf-idf* pour les noms propres, fixée expérimentalement à 25 % :

$$\sigma'(s) = \begin{cases} 0.75\,\sigma(s) & \text{si le mot associé à } s \text{ le plus fréquent dans le texte est un nom propre} \\ \sigma(s) & \text{sinon} \end{cases} \tag{8.2}$$

Ce nouveau critère permet alors de favoriser l'apparition des noms communs parmi les mots-clés.

Indices de confiance

Une autre difficulté propre aux transcriptions automatiques concerne la sélection de mots mal reconnus parmi les mots-clés. Afin de remédier à ce problème, nous utilisons les mesures de confiance calculées à partir des probabilités *a posteriori*, d'après la méthode que nous avons précédemment exposée (*cf.* section 5.4.3). Ces indices nous permettent d'estimer les mots incorrects parmi tous ceux transcrits et de pénaliser ainsi ceux qui sont probablement erronés lors de la détection des mots-clés. Dans la mesure où nous opérons sur des *stems*, nous définissons la mesure de confiance $\mathrm{conf}(s)$ d'un *stem* s sur un texte donné à partir de la moyenne de celles $\mathrm{conf}(s)$ établies pour chaque mot w associé à s dans ce document :

$$\mathrm{conf}(s) = \frac{\sum_{w \in \mathcal{W}_s} \mathrm{NbOcc}(w)\mathrm{conf}(w)}{\sum_{w \in \mathcal{W}_s} \mathrm{NbOcc}(w)} \tag{8.3}$$

où $\mathcal{W}_s$ représente l'ensemble des mots du texte qui ont pour *stem* s et $\mathrm{NbOcc}(w)$ est le nombre d'occurrences de w. Afin d'intégrer les mesures de confiance dans le calcul du *tf-idf*, nous pénalisons les mots risquant d'être mal reconnus, en définissant un nouveau score :

$$\sigma''(s) = \sigma'(s) - \alpha(1 - \mathrm{conf}(s))\sigma'(s) \tag{8.4}$$

où $1 - \mathrm{conf}(s)$ estime la probabilité que le mot soit incorrect et α est un coefficient compris entre 0 et 1 qui amoindrit la pénalité. Nous avons testé le comportement de ce nouveau mode de calcul pour quelques valeurs de α et l'avons fixé à 0,75.

Le score $\sigma''(s)$ est finalement la valeur que nous utilisons pour sélectionner les mots pertinents de chaque section. Nous nous intéressons dans ce qui suit à la manière de construire des requêtes au moteur de recherche, à partir de la liste des mots-clés trouvés.

8.1.2 Formulation de requêtes

Nous utilisons les cinq mots ayant les valeurs $\sigma''(s)$ les plus hautes pour le texte pour interroger le moteur de recherche *Yahoo !*. L'utilisation de l'opérateur ET permet de s'assurer que les documents contiennent plusieurs de ces termes, ce qui renforce potentiellement leur adéquation avec le texte initial. Toutefois, certaines requêtes conduisent à un nombre très réduit de pages ou comportent un mot-clé qui est en réalité peu représentatif de la transcription. Afin de réduire ces effets perturbateurs, nous construisons plusieurs requêtes en réalisant différentes combinaisons de mots-clés et en restreignant les termes utilisés simultanément à trois.

Suivant cette procédure, nous formons nos corpus d'adaptation à partir de l'ensemble des pages Web fournies en réponse par les différentes requêtes. Certaines pages peuvent apparaître plusieurs fois, notamment du fait de l'apparition du même mot-clé dans plusieurs interrogations du moteur de recherche. Nous contrôlons donc ces redondances en n'ajoutant qu'une seule fois une page disponible à un lien HTML donné. La pertinence des résultats tendant à décroître suivant les rangs de réponse fournis par le système de RI et afin de limiter le temps nécessaire à la collecte des données, nous nous restreignons aux 200 meilleures réponses fournies pour chaque requête.

Nous détaillons dorénavant dans la section suivante la manière dont nous procédons pour former des corpus à partir des pages Web parcourues.

8.1.3 Filtrage des résultats

Les données recueillies peuvent être de qualités disparates, d'autant plus que nous ne posons pas de contraintes sur les sites à partir desquels les pages Web ont été obtenues. La constitution d'un corpus utilisable pour la construction de ML nécessite de reformater ces données et de s'assurer qu'elles soient similaires avec le texte original.

Nettoyage de pages Web

Afin d'éliminer toutes les sections inintéressantes pour la constitution de nos corpus, notamment celles correspondant aux menus, aux mentions légales ou à des titres, nous nettoyons les pages HTML en deux temps. Nous parcourons tout d'abord l'arbre DOM (pour *Document Object Model*) représentant la page en ne retenant que les feuilles ayant de bonnes propriétés sur la fréquence des marques de ponctuation et des caractères spéciaux. Nous procédons également à une élimination systématique quand le texte est trop court afin d'écarter les titres ou les références vers d'autres documents. Dans un second temps, les balises HTML restantes sont retirées et les feuilles retenues sont regroupées pour constituer un des textes du corpus d'adaptation.

Filtrage thématique

Afin de limiter dans le corpus la présence d'éléments non pertinents vis-à-vis du document initial, nous vérifions *a posteriori* que les textes retournés partagent

plusieurs mots significatifs avec la transcription initiale. Nous employons à cette fin la mesure de similarité par cosinus, utilisée aussi bien en RI qu'en segmentation thématique (*cf.* section 7.2). Chaque élément du corpus est alors nettoyé puis réduit à un *stem*, de la même manière que le document original. Nous construisons ensuite deux vecteurs constitués des scores $\sigma(s)$ (*cf.* équation (8.1) page 152) des *stems* contenus par les pages retournées et la transcription initiale. Si $\vec{\sigma_t}$ et $\vec{\sigma_d}$ représentent les vecteurs ainsi construits pour chaque couple (t, d) de textes à comparer, la similarité se calcule alors par :

$$sim(t,d) = cos(\vec{\sigma}_t, \vec{\sigma}_d) = \frac{\sum_i \sigma_{t,i}\,\sigma_{d,i}}{\sqrt{\sum_i \sigma_{t,i}^2 \sum_i \sigma_{d,i}^2}} \tag{8.5}$$

où chaque $\sigma_{x,s}$ est le score *tf-idf* du *stem* s dans le texte x. Cette manière de procéder nous permet alors de contrôler que les réponses obtenues comportent les mots-clés utilisés pour les requêtes, mais aussi d'autres présents dans la transcription initiale et qui auraient également pu être pertinents. La constitution des corpus est effectuée en écartant tous les textes t qui ne sont pas assez similaires au document original d, *i.e.*, ceux pour lesquels $sim(t, d)$ est inférieur à un seuil que nous avons fixé à 0,15.

À l'issue de la procédure que nous avons décrite dans cette section, les données obtenues pour chaque document étudié permettent de former des corpus spécifiques, que nous employons dans ce qui suit pour former de nouveaux ML.

8.2 Adaptation du ML

Les corpus recueillis visent à élaborer des ML particuliers à chaque document pour décoder la parole. Les données collectées sont en quantité réduite au regard de l'ensemble d'apprentissage employé pour le ML utilisé par défaut, que nous nommons dorénavant ML généraliste. Pour ces raisons, le ML adapté est élaboré en se basant sur le ML généraliste puis en modifiant certaines probabilités en fonction des séquences de mots trouvées dans les corpus d'adaptation. Dans nos travaux, souhaitant avant tout nous assurer de la validité de notre approche pour obtenir des données caractéristiques à chaque document, nous n'avons considéré qu'une simple méthode d'interpolation linéaire entre le ML généraliste et le ML spécialisé construit à partir du corpus d'adaptation. Le ML adapté ainsi conçu combine les probabilités P_{gen} et P_{spec}, calculées par les deux types précédents de ML, de la manière suivante :

$$P_{\text{adapt}}(w_i|w_1^{i-1}) = \lambda P_{\text{gen}}(w_i|w_{i-3}^{i-1}) + (1-\lambda) P_{\text{spec}}(w_i|w_{i-2}^{i-1}) \tag{8.6}$$

où λ est un coefficient d'interpolation, et où le ML généraliste et le ML spécialisé sont respectivement des ML quadrigrammes et trigrammes.

La mesure de l'apport du ML adapté permet d'évaluer la pertinence des corpus spécifiques mais également de valider l'ensemble de la chaîne de traitements que nous avons proposée au chapitre 6 pour traiter séparément les différentes sections thématiques d'une émission d'information. Notre méthode de segmentation restant encore à améliorer, notamment pour écarter les extraits hétérogènes du point de vue des sujets

traités (*cf.* chapitre 7), nous effectuons ici nos tests sur des sections thématiques sélectionnées manuellement à partir de différentes émissions du corpus de développement d'ESTER. Ces extraits, au nombre de 22, forment mis bout à bout environ une heure de diffusion ; ils ont été choisis de manière à représenter la variabilité des thèmes rencontrés dans l'ensemble du corpus ESTER et celle de la longueur avec laquelle les sujets peuvent être abordés. Parmi cet échantillon, on retrouve ainsi une interview de quinze minutes d'un homme politique à propos du premier tour des élections présidentielles de 2002 en France, des brèves d'une dizaine de secondes sur des faits divers et une chronique sportive relatant différentes disciplines. La procédure de filtrage mise en place pour obtenir des corpus spécifiques à chacune de ces sections nous conduit à retenir en moyenne un tiers des 200 pages retournées, ce qui forme des corpus d'adaptation de l'ordre de 250 000 mots.

Nous donnons dans la suite de cette section deux types de mesure d'évaluation des ML adaptés : la perplexité puis le WER.

8.2.1 Évolution de la perplexité

La perplexité est une métrique souvent utilisée pour juger de la capacité d'un ML pour prédire les mots d'un texte. Nous la déterminons pour chaque section de notre échantillon à partir de la transcription manuelle, et non de celle produite par le système de RAP. Ceci se justifie par le fait que nous cherchons à modéliser au mieux la séquence de mots que nous aurions aimé voir transcrite, et non un texte partiellement incorrect. Nous donnons sur la figure 8.3 la diminution relative de la perplexité obtenue en moyenne pour l'ensemble des segments étudiés, avec la méthode d'adaptation que nous proposons. Un examen de cette courbe montre que la baisse supérieure à 10 % est importante, ce qui est le reflet, en analysant individuellement chaque extrait, d'une amélioration de la perplexité sur toutes les sections thématiques. En fixant le coefficient λ à 0,75 pour toutes les sections, nous observons même une réduction globale de 16,7 %, qui devient égale à 18,0 % quand on choisit séparément λ pour chaque extrait étudié. Ces deux dernières observations nous montrent tout l'intérêt que pourrait avoir l'optimisation de λ pour chaque section.

Nous donnons ci-après quelques résultats mesurant l'impact des choix faits lors de la réalisation de notre prototype. Le tableau 8.2.1 met en évidence l'influence de la prise en compte des mesures de confiance et de celle du filtrage *a posteriori* des pages Web, qui conduisent toutes deux à une baisse plus importante de la perplexité. La figure 8.4 fait voir quant à elle l'avantage qu'il y a à augmenter le nombre d'interrogations du moteur de recherche ; elle montre en outre que le gain en perplexité est principalement observé en élargissant le nombre de pages retenues pour chaque requête au-delà de 50 et que l'amélioration devient moindre à partir de 200 pages.

La baisse importante de la perplexité établit ainsi que les ML adaptés prédisent avec plus de justesse les mots à reconnaître pour chaque section. Nous étudions dans la section suivante l'apport de ces ML sur le gain du WER.

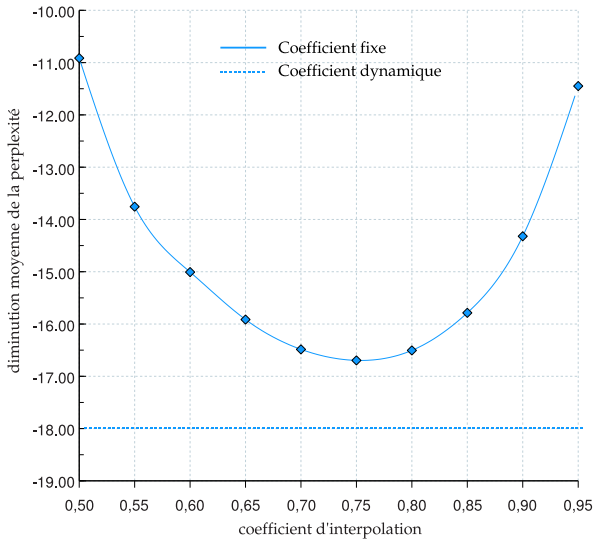

FIG. 8.3: *Diminution relative moyenne de la perplexité en fonction du coefficient d'interpolation du ML généraliste, le coefficient dynamique étant optimisé séparément pour chaque section*

configuration employée	perplexité
avec filtrage et mesure de confiance	-16,7 %
avec filtrage et sans mesure de confiance	-15,5 %
sans filtrage et avec mesure de confiance	-14,5 %

TAB. 8.1: *Baisse moyenne de la perplexité constatée pour plusieurs variantes de la procédure de formation des corpus d'adaptation, en fixant λ à 0,75*

8.2.2 Qualité de la transcription

Afin de mesurer le lien entre la qualité de la transcription et la modification du ML, nous réordonnons les graphes de mots produits par la dernière passe du système IRENE[3] avec le ML adapté construit pour chaque section. La meilleure hypothèse sélectionnée en prenant en compte ce nouveau ML est alors comparée à celle proposée initialement avec le ML généraliste et dont le WER était de 19,1 %. En analysant ainsi un document formé à partir des 22 segments thématiques étudiés, nous avons constaté une diminution absolue du WER de 0,2 %, ce qui est faible au regard de la baisse de la perplexité précédemment observée pour les mêmes textes. En outre, l'examen individuel des sections étudiées montre une hausse du nombre d'erreurs de transcription pour 8 des 22

[3]Sans avoir restreints au préalable aux 1000 meilleurs chemins les graphes produits à l'issue de la première passe (*cf.* section 3.2).

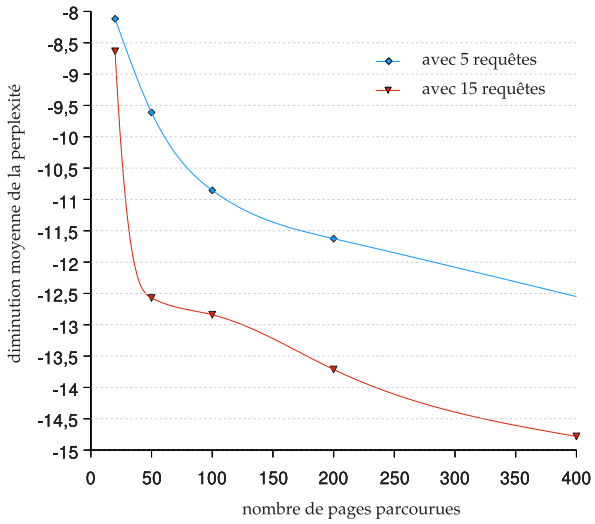

FIG. 8.4: *Variation relative de la perplexité moyenne avec 5 ou 15 requêtes*

segments. Ce manque d'équivalence entre le WER et la perplexité a déjà été observé dans un contexte similaire d'adaptation du ML [Gildea et Hofmann, 1999].

De manière à trouver des raisons qui expliquent cette constatation, nous avons examiné de manière détaillée les modifications induites par le ML. En ce qui concerne les erreurs de transcriptions corrigées, nous avons remarqué comme attendu qu'elles concernaient en majorité les mots liés au thème traité. Nous avons par exemple noté la correction d'erreurs sur le mot composé « *cessez-le-feu* », pour une section relatant de la guerre civile en Côte d'Ivoire, et sur les termes médicaux « *gorge* » et « *bronche* », pour un extrait traitant de la pneumopathie atypique (Fig 8.5). Les erreurs introduites sont au contraire plutôt des fautes réalisées sur des mots ne concernant pas directement le thème abordé par le texte. L'apparition de mots mal reconnus dans le dernier exemple de la figure 8.5 est par exemple dû à la présence fortuite dans le corpus d'adaptation d'un plus grand nombre d'occurrences de la séquence « *ils organisaient* » par rapport à celles de « *il a organisé* ». L'explication principale qui peut donc être donnée sur la différence observée entre la perplexité et le WER est que le ML adapté augmente fortement la probabilité des mots relatifs aux sujets traités et de manière modérée les autres, ce qui au niveau du WER, qui accorde autant d'importance à chaque mot, se manifeste par la correction d'un nombre assez faible d'erreurs.

8.3 Bilan et perspectives

Au vu de l'amélioration certes modérée mais constatée du WER, l'adaptation du ML à partir de corpus issu du Web apparaît comme une approche intéressante. La procédure que nous avons développée pour recueillir des données à partir d'Internet semble permettre d'obtenir des corpus suffisamment pertinents pour engendrer des gains

Erreurs corrigées
REF : SURVEILLENT la ligne de CESSEZ le feu HYP(avant) : SERAIT la ligne de CESSER le feu HYP(après) : SERAIT la ligne de cessez le feu mots-clés : « *réconciliation* », « *côte-d'ivoire* », « *rebelles* », « *abidjan* » et « *ivoirienne* »
REF : DE LA GORGE ET DES BRONCHES HYP(avant) : ** ** L' ACCORD DE BRANCHE HYP(après) : ** la gorge ** des bronches mots-clés : « *grippe* », « *virus* », « *pneumopathie* », « *oms* » et « *rhume* »
Erreurs introduites
REF : y compris ces derniers temps c' est d' ici qu' IL A ORGANISÉ LE montage financier de la fortune accumulée HYP(avant) : y compris ces derniers temps c' est d' ici qu' il a organisé le montage financier de la fortune accumulée HYP(après) : y compris ces derniers temps c' est d' ici qu' ** * ILS ORGANISAIENT montage financier de la fortune accumulée

FIG. 8.5: *Exemples de groupes de souffle modifiés par l'adaptation du ML*

importants de la perplexité. Même si la baisse du WER reste faible, nous pensons qu'elle pourrait être plus importante en concevant un prototype plus abouti pour adapter le ML.

La première modification possible est de remettre en question le choix que nous avons fait d'utiliser uniquement une interpolation linéaire pour nos tests pour spécialiser le ML. L'optimisation des coefficients d'interpolation en fonction de chaque document est ainsi réalisable en cherchant à minimiser la perplexité du ML adapté sur la transcription automatique produite pour l'extrait considéré, ou bien — ce qui permet d'obtenir un plus grand volume de données exempt d'erreurs de transcription — sur le corpus spécifique correspondant. La technique MDI avec des contraintes unigrammes semble de plus être une méthode d'adaptation conduisant à de meilleurs résultats et gagnerait donc à être employée [Chen *et al.*, 2004a].

Une autre évolution assez immédiate de notre système repose sur le repérage de la similarité lexicale entre plusieurs documents en vue de les regrouper. Opérer sur des ensembles de textes permet tout d'abord d'obtenir davantage de mots-clés et d'espérer qu'ils soient meilleurs ; cela évite également d'avoir à consulter à plusieurs reprises le moteur de recherche pour des requêtes similaires.

L'utilisation d'autres modes de représentation des données transcrites peut en outre fournir des pistes supplémentaires. Le ML pourrait notamment être plutôt appliqué sur

les graphes produits à l'issue de la première passe et contenant davantage de mots, dont certains peuvent être liés au sujet abordé par le document. Il est probable qu'un gain plus important au niveau du WER puisse être observé si la richesse au niveau lexicale des graphe de mots exploités est augmenté. En outre, nous ne retenons jusqu'à présent que la meilleure hypothèse pour choisir nos mots-clés. Le recours aux réseaux de confusion, fournissant des probabilités *a posteriori*, permettrait ainsi de prendre en compte un autre problème posé par les erreurs de transcription que celui d'extraire des mots-clés erronés. Certains mots mal reconnus peuvent en effet être remplacé par d'autres et, de manière à surmonter cette difficulté, il apparaît possible de favoriser des mots avec peu d'occurrences dans la transcription mais apparaissant beaucoup plus souvent dans le réseau de confusion avec des hautes valeurs de probabilités *a posteriori*.

L'une des perspectives que nous voyons également se rapporte au critère de sélection des mots-clés. Nous n'avons considéré jusqu'à présent qu'un mode de calcul basé sur la valeur *tf-idf* mais la mesure liée à l'information mutuelle, qui a déjà montré des résultats satisfaisants pour l'adaptation de ML [Chen *et al.*, 2003], constitue une alternative intéressante. Un point qui reste améliorable dans notre méthode concerne la présence des mots rares dans les requêtes formulées au moteur de recherche. En effet, même avec la pénalité que nous avons proposée sur les noms propres, les mots-clés retenus pour quelques sections problématiques, notamment les rubriques sportives contenant de nombreuses entités nommées, conduisent à des corpus de taille très réduite. Une évolution possible consisterait donc à choisir, lorsque le nombre de pages Web retournées est restreint, des termes connectés aux mots-clés, en examinant les mots caractéristiques des quelques textes retournés pour faire une nouvelle recherche, ou en utilisant des connaissances supplémentaires sur les entités nommées.

Ce dernier point nous amène à évoquer l'apport de méthodes issues du TAL, dont notamment celles concernant l'acquisition des relations paradigmatiques. Des techniques d'acquisition sur corpus existent ainsi pour construire des classes sémantiques [Rossignol et Sébillot, 2003], rassemblant des termes possédant une forte similarité au niveau du sens, ou bien encore des classes de mots-clés de mots thématiques [Rossignol et Sébillot, 2006], regroupant des termes employés pour traiter des mêmes domaines. Le premier type de classes peut par exemple être employé pour renforcer la valeur *tf* de termes synonymes, et ainsi leur chance de figurer parmi les mots-clés s'ils sont caractéristiques du document. Le second peut aider à détecter le thème principalement traité dans un document en repérant plusieurs mots d'une même classe thématique. Ces termes *a priori* représentatifs du document peuvent alors être eux aussi favorisés pour apparaître parmi les mots-clés.

Une autre contribution du TAL repose sur le repérage de termes complexes, que l'on peut décrire grossièrement comme des mots composés ou des collocations. Des méthodes de détection de ces termes existent, dont notamment celles basées sur l'identification de patrons comme « NOM PREP (DET) NOM » pour le français [Daille, 2003]. Des groupes de mots, tels que « *président de la république* » ou « *maison blanche* » peuvent ainsi constituer des mots-clés intéressants s'ils sont présents de manière répétée dans le document analysé.

Conclusion

Synthèse

L'objectif fixé dans cette thèse était de proposer de nouvelles méthodes pour améliorer la reconnaissance de la parole au moyen de connaissances sur le langage. Nous avons choisi d'axer notre étude sur deux points : le recours aux informations morpho-syntaxiques et l'adaptation thématique. Nous sommes tout d'abord parti du constat qu'un nombre important d'erreurs commises par notre système de transcription était liées à des fautes d'accord, y compris entre des mots voisins. Nous avons en outre remarqué que les ML employés en RAP sont souvent construits à partir de corpus très généraux, ce qui permet d'assurer une bonne couverture du langage mais a aussi pour effet de bord de niveler les particularités des différents domaines d'une langue.

Notre contribution relative aux POS a consisté à proposer une nouvelle méthode d'intégration de ce type de connaissances en post-traitement d'un système de RAP. L'emploi des informations morpho-syntaxiques n'est pas nouveau mais cette méthode originale nous permet d'obtenir une diminution absolue du WER voisine de 1,0 %, bien supérieure à celle généralement obtenue avec les techniques plus classiques des ML N-classes. Nous avons pour ce faire développé un étiqueteur afin de traiter spécifiquement les transcriptions produites par notre système de RAP et montré qu'il atteignait sur ce type de textes des performances comparables aux meilleurs étiqueteurs. Notre manière de prendre en compte les POS s'intègre en outre parfaitement à différents critères de décodage puisque les trois que nous avons testés conduisent tous à une amélioration similaire de la qualité de la transcription. De par les différents jeux d'étiquettes que nous avons expérimentés, nous montrons également que l'utilisation conjointe d'informations sur la classe grammaticale, et sur le genre et le nombre permet d'augmenter davantage le nombre de mots correctement reconnus. Selon la métrique du LER (pour *lemma error rate*) que nous avons proposée, il apparaît de surcroît que les erreurs corrigées ne se limitent pas aux erreurs d'accord ; la diminution du SER (pour *sentence error rate*) et une analyse des modifications engendrées par la prise en compte des POS indiquent que ce type de connaissance conduit globalement à des groupes de souffle plus « grammaticaux ». Bien que nos expériences se soient limitées essentiellement à la transcription d'émissions francophones, des tests complémentaires que nous avons pu

mener suggèrent que notre approche est intéressante pour des documents relevant de la parole spontanée ou exprimés dans d'autres langues. Notre dernier apport concernant les informations morpho-syntaxiques a enfin été de montrer qu'elles permettaient d'améliorer les mesures de confiance.

Nos travaux ne sont toutefois pas exempts de certaines faiblesses. Le premier d'entre eux est que nous n'opérons que sur des listes d'hypothèses, et non sur des représentations plus riches comme des graphes de mots, ceci étant lié au fait que notre méthode étend chaque hypothèse de mots par l'ensemble des classes catégorielles à laquelle elle peut être associée. Un autre inconvénient est que la correction d'erreurs que nous réalisons repose sur le réordonnancement de listes d'hypothèses, ce qui restreint les mots considérés à ceux proposés par le système de RAP. Nous amenons toutefois un premier élément de réponse avec l'emploi des réseaux de confusion qui autorisent la sélection des suites de mots n'apparaissant pas initialement parmi les listes d'hypothèses. En revanche, notre approche reposant sur l'extension des mots lexicaux par leurs homophones s'est révélée peu concluante puisqu'elle a conduit à une diminution moindre du WER. Enfin, une dernière limite de notre méthode est qu'elle étudie indépendamment chaque groupe de souffle, en ignorant ainsi les dépendances qui peuvent exister entre ces unités de segmentation.

Au cours de la seconde partie, plus prospective, de nos travaux, nous nous sommes intéressé à l'adaptation du ML en fonction des sujets abordés. Notre apport dans ce domaine est d'avoir proposé une chaîne de traitements complète qui nous permet de spécialiser le système de RAP de manière fine, en découpant le document en sections thématiquement homogènes et en construisant des corpus appropriés issus du Web. L'avantage fondamental de notre approche est que notre adaptation est non supervisée puisque nous ne faisons aucune hypothèse sur les thématiques susceptibles d'être traitées dans les documents étudiés et que nous ne nous appuyons pas sur l'existence d'une collection de textes qui serait forcément restreinte à des domaines particuliers. Nous avons élaboré pour ce faire des solutions originales lors de la conception des deux premiers maillons de cette chaîne. Nous avons tout d'abord établi la capacité de la méthode de segmentation thématique que nous avons choisie à être employée sur des transcriptions automatiques. Nous avons en outre proposé plusieurs extensions de ce modèle probabiliste pour mieux prendre en compte les connaissances disponibles dans un document sonore contenant de la parole. Une première, recourant à un réseau de cooccurrences, intègre des informations sur le contexte d'utilisation des mots. Une seconde, reposant sur un nouveau cadre théorique de la segmentation thématique, permet d'inclure les indices acoustiques et le repérage des marqueurs linguistiques lors du processus de détection des frontières entre sections. Nous proposons également un nouveau contexte d'utilisation des modèles N-grammes cachés en établissant leur intérêt pour le repérage des marqueurs linguistiques. En ce qui concerne le deuxième maillon de notre chaîne, consacré à la constitution de corpus spécifiques issus du Web, l'amélioration constatée du WER par des ML adaptés à partir de ces données établit l'intérêt de cette source d'information. Nous avons proposé pour sélectionner les mots représentatifs d'un

document une modification du critère *tf-idf* qui permet de prendre en compte les spécificités de notre contexte applicatif. Nous avons également développé plusieurs méthodes pour utiliser les données recueillies sur le Web comme corpus d'adaptation.

Le prototype développé pour mettre en œuvre notre chaîne de traitements n'est pas encore complet, celui-ci étant réalisé dans un domaine qui demeure encore largement inexploré. Les étapes critiques restant à améliorer sont principalement celles de la segmentation thématique et celles de la sélection des mots-clés, pour lesquelles de nouvelles méthodes sont à développer pour améliorer les résultats. Il reste notamment encore à définir de manière précise le niveau de granularité avec lequel le document à transcrire doit être segmenté. Le critère de découpage qui nous semble important à retenir est celui de l'homogénéité d'une section mais il peut être interprété de différentes manières suivant l'acception de la notion de thème. En ce qui concerne l'adaptation proprement dite du ML, notre méthode reste perfectible mais d'autres techniques plus performantes semblent exister [Chen *et al.*, 2004b].

Perspectives

Outre les perspectives déjà mentionnées en conclusion des chapitres 7 et 8 sur l'amélioration de notre chaîne de traitements pour adapter thématiquement le ML, nos travaux font apparaître d'autres extensions possibles.

Tout d'abord, notre méthode pour prendre en compte les POS en RAP se limite à examiner séparément chaque groupe de souffle alors que des liens syntaxiques peuvent exister entre des unités voisines. Afin de prendre en compte des dépendances à plus longue distance lors de l'utilisation des informations morpho-syntaxiques, une segmentation linguistique paraît plus appropriée que celle en groupes de souffle que nous avons considérée jusqu'à maintenant. La détection des fins de phrase dans une transcription reste encore un problème difficile [Liu *et al.*, 2004] mais il est possible qu'un découpage même imparfait conduise à une amélioration supplémentaire du WER.

Nous voyons également deux autres pistes, issues du domaine du TAL. Une première repose sur la détection des anaphores dans la transcription. En repérant notamment à quoi font référence les pronoms personnels, il devient ainsi possible de discriminer « *il* » et « *ils* », ou « *elle* » et « *elles* ». La résolution automatique d'anaphores demeure cependant une tâche ardue, y compris pour des textes de la langue écrite. Une seconde perspective repose sur une analyse de surface des transcriptions. La segmentation en *chunks* peut ainsi mettre au jour des relations entre un groupe nominal et un verbe — même s'ils sont distants — ou bien permettre de détecter certains groupes de souffle non grammaticaux. Le principal obstacle à l'emploi de cette technique est lié aux « écarts » de la transcription automatique par rapport au langage conventionnel. Ce type d'analyseur a toutefois déjà été employé sur de la parole spontanée avec de bonnes performances [Zechner et Waibel, 1998].

Une extension de nos travaux portant sur l'adaptation consiste à étendre la taille du vocabulaire du système de RAP. Nous n'avons jusqu'à présent considéré qu'une modification du ML mais la prise en compte du thème traité dans un document verrait

son intérêt augmenté en ajoutant au lexique des mots particuliers à ce domaine. Cette propriété est particulièrement pertinente pour transcrire des émissions d'information, où figurent de nombreuses entités nommées qui, à part les plus courantes, ne font pas partie du dictionnaire.

Enfin, les méthodes que nous avons développées dans la seconde partie de nos travaux s'inscrivent naturellement dans un cadre applicatif qui est celui de la structuration de flux multimédia. Nos techniques de segmentation thématique et de détection de mots-clés permettent en effet de découper automatiquement une émission selon les thèmes abordés, en fournissant une brève description par quelques mots jugés caractéristiques. Ces indications apportées sur le contenu sémantique peuvent être employées pour naviguer de manière intelligente dans un grand ensemble de documents sonores contenant de la parole. Parmi les applications possibles, citons le suivi d'un sujet ou la recherche d'une information parmi le programme diffusé par une chaîne de radio ou de télévision.

Bibliographie

Abney S. (1996). *Corpus-Based Methods in Language and Speech*, chapitre Part-of-Speech Tagging and Partial Parsing, pages 118–136. Kluwer Academic Publishers, Dordrecht, Pays-Bas.

Adda G., Mariani J., Paroubek P., Rajman M. et Lecomte J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. *Dans Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France.

Adda-Decker M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *Dans Actes des 26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

Akita Y., Nemoto Y. et Kawahara T. (2007). PLSA-Based Topic Detection in Meetings for Adaptation of Lexicon and Language Model. *Dans Proc. of the 10th European Conference on Speech Communication and Technology (Eurospeech)*, Anvers, Belgique.

Alain P. et Boëffard O. (2006). Algorithme de recherche d'un rang de prédiction. Application à l'évaluation de modèles de langage. *Dans Actes des 26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

Allan J., Carbonell J., Doddington G., Yamron J. et Yang Y. (1998). Topic Detection and Tracking Pilot Study Final Report. *Dans Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginie, États-Unis.

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

Allauzen A. et Gauvain J.-L. (2003). Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés. *Traitement Automatique des Langues (TAL)*, 44(1):11–31.

Allauzen A. et Gauvain J.-L. (2005). Open Vocabulary ASR for Audiovisual Document Indexation. *Dans Proc. of the IEEE International Conference on Acous-*

tics, Speech, and Signal Processing (ICASSP), volume 1, Philadelphie, Pennsylvanie, États-Unis.

Allen J. (1976). Synthesis of Speech from Unrestricted Text. *Proc. of the IEEE*, 64(4):433–442.

Allen J. B. (1994). How do Humans Process and Recognize Speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577.

Antoine J.-Y. et Genthial D. (1999). Méthodes hybrides issues du TALN et du TAL parlé : état des lieux et perspectives. *Dans Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France.

Antoine J.-Y. et Goulian J. (2001). Word Order Variations and Spoken Man-Machine Dialogue in French: a Corpus Analysis on the ATIS Domain. *Dans Proc. of Corpus Linguistics*, Lancaster, Royaume-Uni.

Armstrong S., Bouillon P. et Robert G. (1995). Tools for Part-of-Speech Tagging. Rapport technique, ISSCO, Genève, Suisse.

Asadi A., Schwartz R. et Makhoul J. (1990). Automatic Detection of New words in a Large Vocabulary Continuous Speech Recognition System. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Albuquerque, Nouveau Mexique, États-Unis.

Bacchiani M. et Roark B. (2003). Unsupervised Language Model Adaptation. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Hong Kong, Chine.

Banko M. et Brill E. (2001). Scaling to Very Large Corpora for Natural Language Disambiguation. *Dans Proc. of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, Toulouse, France.

Baum L. E., Petrie T., Soules G. et Weiss N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41:164–171.

Béchet F., Nasr A., Spriet T. et de Mori R. (1999). Modèles de langage à portée variable : application au traitement des homophones. *Dans Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France.

Beeferman D., Berger A. et Lafferty J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210.

Bellegarda J. R. (1998). A Multispan Language Modeling Framework for Large Vocabulary Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(5):456–467.

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

Bellegarda J. R. (2004). Statistical Language Model Adaptation: Review and Perspectives. *Speech Communication*, 42(1):93–108.

Benzitoun C., Campione E., Deulofeu J., Henry S., Sabio F., Teston S., Valli A. et Véronis J. (2004). L'analyse syntaxique de l'oral : problèmes et méthode. *Dans Actes de la journée d'étude de l'ATALA sur l'annotation syntaxique de corpus*, Paris, France.

Berger A. et Miller R. (1998). Just-in-Time Language Modelling. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Seattle, Washington, États-Unis.

Berger A. et Printz H. (1998). Recognition Performance of a Large-Scale Dependency-Grammar Language Model. *Dans Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*, volume 6, Sydney, Australie.

Bestgen Y. (2006). Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

Bigi B., De Mori R., El-Bèze M. et Spriet T. (1998). Detecting Topic Shift Using a Cache Memory. *Dans Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*, volume 6, Sydney, Australie.

Bigi B., De Mori R. et Spriet T. (2000). Reconnaissance thématique à partir de textes dictés et adaptation dynamique de modèles de langage thématiques. *Dans Actes des 23èmes Journées d'Études sur la Parole (JEP)*, Aussois, France.

Bigi B., Huang Y. et De Mori R. (2004). Vocabulary and Language Model Adaptation Using Information Retrieval. *Dans Proc. of the 8th International Conference on Spoken Language Processing (ICSLP)*, volume 2, île de Jeju, Corée du Sud.

Bimbot F., El-Bèze M. et Jardinot M. (1997). An Alternative Scheme for Perplexity Estimation. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Munich, Allemagne.

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

Blanche-Benveniste C. (1990). *Le français parlé : études grammaticales*. Paris : Éditions du CNRS.

Blei D. M. et Moreno P. J. (2001). Topic Segmentation with an Aspect Hidden Markov Model. *Dans Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, La Nouvelle-Orléans, Louisiane, États-Unis.

Blei D. M., Ng A. Y. et Jordan M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bouaud J., Habert B., Nazarenko A. et Zweigenbaum P. (1997). Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. *Dans Actes des journées francophones d'ingénierie des connaissances (IC)*, Roscoff, France.

Boufaden N., Lapalme G. et Bengio Y. (2002). Segmentation en thèmes de conversations téléphoniques:traitement en amont pour l'extraction d'information. *Dans Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, Nancy, France.

Bourlard H. et Dupont S. (1996). A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. *Dans Proc. of 4th International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie, Pennsylvanie, États-Unis.

Brants T. (2000). TnT - A Statistical Part-of-Speech Tagger. *Dans Proc. of the 6th Conference on Applied Natural Language Processing (ANLP)*, Seattle, Washington, États-Unis.

Brill E. (1992). A Simple Rule-Based Part of Speech Tagger. *Dans Proc. of the 3rd Conference on Applied Natural Language Processing (ANLP)*, Trento, Italie. Association for Computational Linguistics.

Brill E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *Dans Proc. of the 12th National Conference on Artificial Intelligence*, volume 1, Seattle, Washington, États-Unis. American Association for Artificial Intelligence.

Brill E., Florian R., Henderson J. C. et Mangu L. (1998). Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling? *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, volume 1, Montréal, Canada.

Brown G. et Yule G. (1983). *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Cambridge University Press.

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, Nancy, France.

Bulyko I., Ostendorf M., Siu M., Ng T., Stolcke A. et Çetin Ö. (2007). Web Resources for Language Modeling in Conversational Speech Recognition. *ACM Transactions on Speech and Language Processing (TSLP)*, 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 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, volume 2, Edmonton, Canada.

Caillet M., Pessiot J.-F., Amini M.-R. et Gallinari P. (2004). Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts. *Dans Proc. of recherche d'information assistée par ordinateur (RIAO)*, Avignon, France.

Campione E. (2001). *Étiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*. Thèse de doctorat, Université de Provence, Aix-en-Provence, France.

Campione E. et Véronis J. (2002). Étude des relations entre pauses et ponctuations pour la synthèse de la parole à partir de texte. *Dans Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Nancy, France.

Campione E. et Véronis J. (2004). Pauses et hésitations en français spontané. *Dans Actes des 25èmes Journées d'Études sur la Parole (JEP)*, Fès, Maroc.

Campione E., Véronis J. et Deulofeu J. (2005). *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, chapitre 3. The French corpus, pages 111–133. Amsterdam: John Benjamins.

Candea M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'ńhésitationż en français oral spontané. Étude sur un corpus de récits en classe de français*. Thèse de doctorat, Université Paris 3, Paris, France.

Carlson B. A. (1996). Unsupervised Topic Clustering of SWITCHBOARD Speech Messages. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Atlanta, Géorgie, États-Unis.

Çarki K., Geutner P. et Schultz T. (2000). Turkish LVCSR: Towards Better Speech Recognition for Agglutinative Languages. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, Istanbul, Turquie.

Charniak E. (2000). A Maximum-Entropy-Inspired Parser. *Dans Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, Washington, États-Unis.

CHARNIAK E. (2001). Immediate-Head Parsing for Language Models. *Dans Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France.

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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Hong Kong, Chine.

CHEN L., GAUVAIN J.-L., LAMEL L. et ADDA G. (2004a). Dynamic Language Modeling for Broadcast News. *Dans Proc. of the 8th International Conference on Spoken Language Processing (ICSLP)*, île de Jeju, Corée du Sud.

CHEN L., GAUVAIN J.-L., LAMEL L., ADDA G. et ADDA-DECKER M. (2001). Using Information Retrieval Methods for Language Model Adaptation. *Dans Proc. of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Danemark.

CHEN L., LIU Y., HARPER M. P. et SHRIBERG E. (2004b). Multimodal Model Integration for Sentence Unit Detection. *Dans Proc. of Int. Conf. Multimodal Interfaces (ICMI)*, State College, Pennsylvanie, États-Unis.

CHEN S. F. et GOODMAN J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Rapport technique, Harvard University, Cambridge, Massachusetts, États-Unis.

CHOI F. Y. Y. (2000). Advances in Domain Independent Linear Text Segmentation. *Dans Proc. of the 6th Applied Natural Language Conference (ANLP) and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, Washington, États-Unis.

CHOI F. Y. Y., WIEMER-HASTINGS P. et MOORE J. (2001). Latent Semantic Analysis for Text Segmentation. *Dans Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pittsburgh, Pennsylvanie, États-Unis.

CHOW Y.-L. et ROUKOS S. (1989). Speech Understanding Using a Unification Grammar. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Glasgow, Royaume-Uni.

CHOW Y.-L. et SCHWARTZ R. (1989). The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses. *Dans Proc. of the DARPA Speech and Natural Language Workshop*, Philadelphie, Pennsylvanie, États-Unis.

CLARKSON P. et ROBINSON T. (1999). Towards Improved Language Model Evaluation Measures. *Dans Proc. of the 6th European Conference on Speech Communication and Technology (Eurospeech)*, volume 5, Budapest, Hongrie.

Clarkson P. R. et Robinson A. J. (1997). Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Munich, Allemagne.

Daille B. (2003). Conceptual structuring through term variations. *Dans Proc. of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japon.

Deligne S. et Bimbot F. (1995). Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, Michigan, États-Unis.

Deligne S. et Sakisaga Y. (1998). Learning a Syntagmatic and Paradigmatic Structure from Language Data with a Bi-Multigram Model. *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, volume 1, Montréal, Canada.

Demetriou G., Atwell E. et Souter C. (1997). Large-Scale Lexical Semantics for Speech Recognition Support. *Dans Proc. of the 5th European Conference on Speech, Communication, Technology (Eurospeech)*, Rhodes, Grèce.

Deshmukh N., Ganapathiraju A. et Picone J. (1999). Hierarchical Search for Large Vocabulary Conversational Speech Recognition. *IEEE Signal Processing Magazine*, 16(5):84–107.

Dharanipragada S., Franz M., McCarley J. S., Roukos S. et Ward T. (1999). Story Segmentation and Topic Detection in the Broadcast News Domain. *Dans Proc. of the DARPA Broadcast News Workshop*, Herndon, Virginie, États-Unis.

El-Bèze M. et Derouault A.-M. (1990). A Morphological Model for Large Vocabulary Speech Recognition. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Albuquerque, Nouveau Mexique, États-Unis.

Équipe DELIC (2004). Présentation du corpus de référence du français parlé. *Recherches sur le français parlé*, 18.

Farhat A., Isabelle J.-F. et O'Shaughnessy D. (1996). Clustering Words for Statistical Language Models Based on Contextual Word Similarity. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Atlanta, Géorgie, États-Unis.

Federico M. (1996). Bayesian Estimation Methods for N-Gram Language Model Adaptation. *Dans Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie, Pennsylvanie, États-Unis.

FEDERICO M. (1999). Efficient Language Model Adaptation through MDI Estimation. *Dans Proc. of the 6th European Conference on Speech, Communication, Technology (Eurospeech)*, volume 4, Budapest, Hongrie.

FELLBAUM C., éditeur (1998). *WordNet: An Electronic Lexical Database*. MIT press, Cambridge, Massachusetts, États-Unis.

FERNÁNDEZ S., SANJUAN É. et TORRES-MORENO J. M. (2007). Énergie textuelle de mémoires associatives. *Dans Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, Toulouse, France.

FERRET O. (1998). *ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage*. Thèse de doctorat, Université Paris 11, Orsay, France.

FERRET O. et GRAU B. (2001). Utiliser des corpus pour amorcer une analyse thématique. *Traitement Automatique des Langues (TAL)*, 42(2):517–545.

FLORIAN R. et YAROWSKY D. (1999). Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation. *Dans Proc. of 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, Maryland, États-Unis.

GALE W. A. et SAMPSON G. (1995). Good-Turing Frequency Estimation without Tears. *Journal of Quantitative Linguistics*, 2:217–237.

GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. et JING H. (2003). Discourse Segmentation of Multi-Party Conversation. *Dans Proc. of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, volume 1, Sapporo, Japon.

GARSIDE R. (1995). *Spoken English on Computer: Transcription, Mark-up and Application*, chapitre Grammatical Tagging of the Spoken Part of the British National Corpus: A Progress Report, pages 161–167. Harlow: Longman.

GAUVAIN J.-L., ADDA G., ADDA-DECKER M., ALLAUZEN A., GENDNER V., LAMEL L. et SCHWENK H. (2005). Where are we in Transcribing French Broadcast News? *Dans Proc. of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, Lisbonne, Portugal.

GAUVAIN J.-L., LAMEL L., ADDA G. et ADDA-DECKER M. (1994). The LIMSI Continuous Speech Dictation System. *Dans Proc. of the ARPA Workshop on Human Language Technology (HLT)*, Plainsboro, New Jersey, États-Unis.

GENDNER V. et ADDA-DECKER M. (2002). Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques. *Dans Actes des 24èmes Journées d'Études sur la Parole (JEP)*, Nancy, France.

GEORGESCUL M., CLARK A. et ARMSTRONG S. (2006a). An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms. *Dans Proc. of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australie.

GEORGESCUL M., CLARK A. et ARMSTRONG S. (2006b). Word Distributions for Thematic Segmentation in a Support Vector Machin Approach. *Dans Proc. of the 10th Conference on Computational Natural Language Learning (CoNLL)*, New York, New York, États-Unis.

GEORGESCUL M., CLARK A. et ARMSTRONG S. (2007). Exploiting Structural Meeting-Specific Features for Topic Segmentation. *Dans Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, Toulouse, France.

GEUTNER P. (1996). Introducing Linguistic Constraints into Statistical Language Modeling. *Dans Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphie, Pennsylvanie, États-Unis.

GHADESSY M., éditeur (1995). *Thematic Development in English Texts*. Pinter Pub Ltd.

GILDEA D. et HOFMANN T. (1999). Topic-Based Language Models Using EM. *Dans Proc. of the 6th European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hongrie.

GILLETT J. et WARD W. (1998). A Language Model Combining Trigrams and Stochastic Context-Free Grammars. *Dans Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*, volume 6, Sydney, Australie.

GILLICK L. et COX S. J. (1989). Some Statistical Issues in the Comparison of Speech Recognition Algorithms. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Glasgow, Royaume-Uni.

GIMÉNEZ J. et MÀRQUEZ L. (2004). SVMTool: A General POS Tagger Generator Based on Support Vector Machines. *Dans Proc. of the 4th International Conference on Language Resources and Evaluation (LREC)*, volume 1, Lisbonne, Portugal.

GOOD I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40(3-4):237–264.

GOODMAN J. T. (2001). A Bit of Progress in Language Modeling, Extended Version. Rapport technique, Microsoft Research, Redmond, Washington, États-Unis.

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)*, Fès, Maroc.

GRAVIER G., YVON F., JACOB B. et BIMBOT F. (2001). Integrating Contextual Phonological Rules in a Large Vocabulary Decoder. *Dans Proc of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Danemark.

GRETTER R. et RICCARDI G. (2001). On-Line Learning of Language Models with Word Error Probability Distributions. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Salt Lake City, Utah, États-Unis.

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 International Conference on Computational Linguistics (COLING)*, volume 1, Taipei, Taiwan.

HALL K. et JOHNSON M. (2004). Attention Shifting for Parsing Speech. *Dans Proc. of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, Barcelone, Espagne.

HALLIDAY M. A. K. et HASAN R. (1976). *Cohesion in English*. Longman Group, London.

HAMAKER J. S. (1999). Towards Building a Better Language Model for SWITCHBOARD: the POS Tagging Task. *Dans Proc. of the International Conference on Information Intelligence and Systems (ICIIS)*, Rockville, Maryland, États-Unis.

HARPER M. P. et HELZERMAN R. A. (1995). Extensions to Constraint Dependency Parsing for Spoken Language Processing. *Computer Speech and Language*, pages 187–234.

HARPER M. P., JAMIESON L. H., MITCHELL C. D., YING G., POTISUK S., SRINIVASAN P. N., CHEN R., ZOLTOWSKI C. B., MCPHETERS L. L., PELLOM B. et HELZERMAN R. A. (1994). Integrating Language Models with Speech Recognition. *Dans Proc. of the AAAI94 Workshop on the Integration of Natural Language and Speech Processing*, Seattle, Washington, États-Unis.

HARPER M. P., JOHNSON M. T., JAMIESON L. H., HOCKEMA S. A. et WHITE C. M. (1999). Interfacing a CDG Parser with an HMM Word Recognizer Using Word Graphs. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Phoenix, Arizona, États-Unis.

HART P., NILSSON N. et RAPHAEL B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics (SSC4)*, 2:100–107.

HAUENSTEIN A. et WEBER H. (1994). An Investigation of Tightly-Coupled Time-Synchronous Speech Language Understanding Using a Unification Grammar. *Dans Proc. of the 12th National Conference on Artificial Intelligence Workshop on the Integration of Natural Language and Speech Processing*, Seattle, Washington, États-Unis.

Hearst M. A. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Heeman P. A. (1999). POS Tags and Decision Trees for Language Modeling. *Dans Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Maryland, États-Unis.

Heeman P. A. et Allen J. F. (1999). Speech Repairs, Intonational Phrases, and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. *Computational Linguistics*, 25(4):527–571.

Heinonen O. (1998). Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, Montréal, Canada.

Henry S. (2002). Quelles répétitions à l'oral ? Esquisse d'une typologie. *Dans Actes des 2èmes Journées de Linguistique de Corpus*, Lorient, France.

Henry S. et Pallaud B. (2003). Word Fragments and Repeats in Spontaneous Spoken French. *Dans Proc. of Disfluency in Spontaneous Speech Workshop (DISS)*, Göteborg, Suède.

Huet S., Gravier G. et Sébillot P. (2006a). Are Morphosyntactic Taggers Suitable to Improve Automatic Transcription? *Dans Proc. of the 9th International Conference on Text, Speech and Dialogue (TSD)*, volume 4188/2006 de *Lecture Notes in Computer Science*, Brno, République tchèque. Springer.

Huet S., Gravier G. et Sébillot P. (2007). Morphosyntactic Processing of N-Best Lists for Improved Recognition and Confidence Measure Computation. *Dans Proc. of the 10th European Conference on Speech Communication and Technology (Eurospeech)*, Anvers, Belgique.

Huet S., Lecorvé G., Gravier G. et Sébillot P. (2008). *Multimodal Processing and Interaction: Audio, Video, Text.*, volume 33 de *"Multimedia Systems and Applications"*, chapitre Toward the Integration of Natural Language Processing and Automatic Speech Recognition: Using Morpho-syntax and Pragmatics for Transcription. Springer.

Huet S., Sébillot P. et Gravier G. (2006b). Introduction de connaissances linguistiques en reconnaissance de la parole : un état de l'art. Rapport de recherche 5917, INRIA.

Hull D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. *Dans Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, Pennsylvanie, États-Unis.

Hurault-Plantet M., Jardino M. et Berthelin J.-B. (2006). Ajustement des frontières de segments thématiques détectés automatiquement. *Dans Actes du 2ème défi fouilles de textes (DEFT)*, Fribourg, Suisse.

Ide I., Mo H., Katayama N. et Satoh S. (2003). Topic-Based Inter-Video Structuring of a Large Scale News Video Corpus. *Dans Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, Baltimore, Maryland, États-Unis.

Ide N. et Véronis J. (1994). MULTEX: Multilingual Text Tools and Corpora. *Dans Proc. of the 15th International Conference on Computational Linguistics (COLING)*, volume 1, Kyoto, Japon.

Ide N. et Véronis J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):2–40.

Isotani R. et Matsunaga S. (1994). Speech Recognition Using a Stochastic Language Model Integrating Local and Global Constraints. *Dans Proc. of the Workshop on Human Language Technology*, Plainsboro, New Jersey, États-Unis.

Iyer R. et Ostendorf M. (1999). Modeling Long Distance Dependence in Language: Topic Mixtures *versus* Dynamic Cache Models. *IEEE Transactions on Speech and Audio Processing*, 7(1):30–39.

Jardino M. (1996). Multilingual Stochastic N-Gram Class Language Models. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Atlanta, Géorgie, États-Unis.

Jelinek F. (1990). *Readings in Speech Recognition*, chapitre Self-Organized Language Modeling for Speech Recognition, pages 450–506. Morgan Kaufmann Publishers.

Jelinek F. (1998). *Statistical Methods for Speech Recognition*. The MIT Press.

Jelinek F. et Lafferty J. D. (1991). Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars. *Computational Linguistics*, 17(3):315–323.

Ji X. et Zha H. (2003). Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. *Dans Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada.

Jiang H. (2005). Confidence Measures for Speech Recognition: A Survey. *Speech Communication*, 45:455–470.

Jobbins A. C. et Evett L. J. (1998). Text Segmentation Using Reiteration and Collocation. *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, Montréal, Canada.

Juneja A. (2004). *Speech Recognition Based on Phonetic Features and Acoustic Landmarks*. Thèse de doctorat, University of Maryland, College Park, Maryland, États-Unis.

Jurafsky D. et Martin J. H. (2008). *Speech and Natural Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2ème édition. À paraître (disponible partiellement sur `http://www.cs.colorado.edu/~martin/slp2.html`).

Jurafsky D., Wooters C., Segal J., Stolcke A., Fosler E., Tajchman G. et Morgan N. (1995). Using a Stochastic Context-Free Grammar as a Language Model for Speech Recognition. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Detroit, Michigan, États-Unis.

Kan M.-Y., Klavans J. L. et McKeown K. R. (1998). Linear Segmentation and Segment Significance. *Dans Proc. of the 6th Workshop on Very Large Corpora (WVLC)*, Montréal, Canada.

Katz S. M. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

Kaufmann S. (1999). Cohesion and Collocation: Using Context Vectors in Text Segmentation. *Dans Proc. of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, College Park, Maryland, États-Unis.

Kehagias A., Pavlina F. et Petridis V. (2003). Linear Text Segmentation Using a Dynamic Programming Algorithm. *Dans Proc. of the 10th Conference European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Hongrie.

Kemp T. et Waibel A. (1998). Reducing the OOV Rate in Broadcast News Speech Recognition. *Dans Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australie.

Khudanpur S. et Wu J. (1999). A Maximum Entropy Language Model to Integrate N-Grams and Topic Dependencies for Conversational Speech Recognition. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, Arizona, États-Unis.

Kilgarriff A. et Grefenstette G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

Kim W. (2004). *Language Model Adaptation for ASR and Statistical MT*. Thèse de doctorat, Johns Hopkins University, Baltimore, Maryland, États-Unis.

Kirchhoff K., Bilmes J. et Duh K. (2007). Factored Language Model Tutorial. Rapport technique, University of Washington, Seattle, Washington, États-Unis.

Kita K., Kawabata T. et Saito H. (1989). HMM Continuous Speech Recognition Using Predictive LR Parsing. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Glasgow, Royaume-Uni.

Klakow D. (2000). Selecting Articles from the Language Model Training Corpus. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, Istanbul, Turquie.

Kneser R. (1996). Statistical Language Modeling Using a Variable Context Length. *Dans Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie, Pennsylvanie, États-Unis.

Kneser R. et Ney H. (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. *Dans Proc. of the 3rd European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, Berlin, Allemagne.

Kneser R. et Ney H. (1995). Improved Backing-Off for M-Gram Language Modeling. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Detroit, Michigan, États-Unis.

Kneser R. et Steinbiss V. (1993). On the Dynamic Adaptation of Stochastic Language Models. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, Minneapolis, Minnesota, États-Unis.

Koehn P., Och F. J. et Marcu D. (2003). Statistical Phrase-Based Translation. *Dans roc. of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, volume 1, Edmonton, Canada.

Kozima H. (1993). Text Segmentation Based on Similarity between Words. *Dans Proc. of the 31st annual meeting on Association for Computational Linguistics (ACL)*, Columbus, Ohio, États-Unis.

Kuhn R. et De Mori R. (1990). A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6): 570–583.

Kuo H.-K. J. et Reichl W. (1999). Phrase-Based Language Models for Speech Recognition. *Dans Proc. of the 6th European Conference on Speech Comunication and Technology (Eurospeech)*, Budapest, Hongrie.

Lafferty J., Sleator D. et Temperley D. (1992). Grammatical Trigrams: A Probabilistic Link Grammar. *Dans Proc. of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, Massachusetts, États-Unis.

Lane I. R., Kawahara T., Matsui T. et Nakamura S. (2005). Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching. *IEICE Transactions 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.

Lapata M. et Keller F. (2005). Web-Based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–31.

Lavecchia C., Smaïli K. et Haton J.-P. (2006). How to Handle Gender and Number Agreement in Statistical Language Models? *Dans Proc.of the 9th International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, Pennsylvanie, États-Unis.

Lecomte J. (1997). Codage Multext pour GRACE/MULTITAG. Critères d'assignation des étiquettes morpho-syntaxiques. Documentation de GRACE.

Lecorvé G. (2007). Adaptation thématique d'un système de transcription automatique de la parole. Mémoire de master recherche, INSA de Rennes, Rennes, France.

Leggetter C. J. et Woodland P. C. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Model. *Computer Speech and Language*, 9(2):171–185.

Levenshtein V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Li H. et Yamanishi K. (2000). Topic Analysis Using a Finite Mixture Model. *Dans Proc. of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, Hong Kong, Chine.

Lin D. et Pantel P. (2001). Induction of Semantic Classes from Natural Language Text. *Dans Proc. of the 7th international conference on Knowledge discovery and data mining (SIGKDD(*, San Francisco, Californie, États-Unis.

Linares D., Benedí J.-M. et Sánchez J.-A. (2004). A Hybrid Language Model Based on a Combination of N-Grams and Stochastic Context-Free Grammars. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):113–127.

Liu S. A. (1995). *Landmark Detection for Distinctive Feature-Based Speech Recognition.* Thèse de doctorat, Massachusetts Institute of Technology, Cambridge, Massachusetts, États-Unis.

Liu Y., Stolcke A., Harper M. P. et Shriberg E. (2004). Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech. *Dans Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelone, Espagne.

MAHAJAN M., BEEFERMAN D. et HUANG X. D. (1999). Improved Topic-Dependent Language Modeling Using Informationretrieval Techniques. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Phoenix, Arizona, États-Unis.

MALTESE G. et MANCINI F. (1992). An Automatic Technique to Include Grammatical and Morphological Information in a Trigram-Based Statistical Language Model. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, San Francisco, Californie, États-Unis.

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.

MARCUS M. P., SANTORINI B. et MARCINKIEWICZ M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19(2): 313–330.

MARTIN S. C., LIERMANN J. et NEY H. (1997). Adaptive Topic Dependent Language Modelling Using Word-Based Varigrams. *Dans Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Grèce.

MELIS L. (2000). Le français parlé et le français écrit, une opposition à géométrie variable. *Romaneske*, 25(3):56–66.

MENDES A., AMARO R. et BACELAR DO NASCIMENTO M. F. (2003). Reusing Available Resources for Tagging a Spoken Portuguese Corpus. *Dans Proc. of the Workshop on Tagging and Shallow Processing of Portuguese (TASHA)*, Lisbonne, Portugal.

MERIALDO B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20(2):155–171.

MORARU D. et GRAVIER G. (2006). Décodage avec ancrage pour la reconnaissance automatique de la parole. *Dans Actes des 26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

MORENO A. et GUIRAO J. M. (2003). Tagging a Spontaneous Speech Corpus of Spanish. *Dans Proc. of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgarie.

MORRIS J. et HIRST G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

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 International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, États-Unis.

NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, 41(2):523–547.

NASR A., ESTÈVE Y., BÉCHET F., SPRIET T. et DE MORI R. (1999). A Language Model Combining N-Grams and Stochastic Finite State Automata. *Dans Proc. of the 6th European Conference on Speech Communication and Technology (Eurospeech)*, volume 5, Budapest, Hongrie.

NEY H. (1991). Dynamic Programming Parsing for Context-Free Grammars in Continuous Speech Recognition. *IEEE Transactions on Signal Processing*, 39(2):336–340.

NIESLER T. R., WHITTAKER E. W. D. et WOODLAND P. C. (1998). Comparison of Part-of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, Washington, États-Unis.

NIESLER T. R. et WOODLAND P. C. (1996a). Combination of Word-Based and Category-Based Language Models. *Dans Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie, Pennsylvanie, États-Unis.

NIESLER T. R. et WOODLAND P. C. (1996b). A Variable-Length Category-Based N-Gram Language Model. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Atlanta, Géorgie, États-Unis.

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 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Danemark.

NIVRE J. (2000). Sparse Data and Smoothing in Statistical Part-of-Speech Tagging. *Journal of Quantitative Linguistics*, 7:1–17.

NIVRE J. et GRÖNQVIST L. (2001). Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics*, 6(1):47–78.

OKUMURA M. et HONDA T. (1994). Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. *Dans Proc. of the 15th International Conference on Computational Linguistics (COLING)*, volume 2, Kyoto, Japon.

ORTMANNS S., NEY H. et AUBERT X. (1997). A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. *Computer, Speech and Language*, 11(1): 43–72.

PALLAUD B. et HENRY S. (2004). Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. *Dans Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Louvain-la-Neuve, Belgique.

Pallett D. S. (2003). A Look at NIST's Benchmark ASR Tests: Past, Present, and Future. *Dans Proc. of the IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, îles Vierges, États-Unis.

Panunzi A., Picchi E. et Moneglia M. (2004). Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian. *Dans Proc. of the 4th International Conference on Language Resources and Evaluation (LREC)*, volume 2, Lisbonne, Portugal.

Paroubek P., Robba I., Vilnat A. et Ayache C. (2006). Data, Annotations and Measures in EASY, the Evaluation Campaign for Parsers of French. *Dans Proc. of the 5th International Conference on Language Resources and Evaluation (LREC)*, Gênes, Italie.

Passonneau R. J. et Litman D. J. (1993). Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. *Dans Proc. of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, Ohio.

Passonneau R. J. et Litman D. J. (1997). Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, 23(1):103–139.

Peng F. et Schuurmans D. (2001). A Simple Closed-Class/Open-Class Factorization for Language Modeling. *Dans Proc. of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, Tokyo, Japon.

Pevzner L. et Hearst M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36.

Polguère A. (2003). *Lexicologie et sémantique lexicale : notions fondamentales.* Les Presses de l'Université de Montréal.

Porter M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

Quiniou S., Anquetil É. et Carbonnel S. (2005). Statistical Language Models for On-line Handwritten Sentence Recognition. *Dans Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, Séoul, Corée du Sud.

Quinlan J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Rabiner L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–285.

Rastier F. (1995). *L'analyse thématique des données textuelles*, chapitre La sémantique des thèmes ou le voyage sentimental, pages 223–249. Didier, Paris.

Ratnaparkhi A. (1996). A Maximum Entropy Part-of-Speech Tagger. *Dans Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphie, Pennsylvanie, États-Unis.

RATNAPARKHI A. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Rapport technique, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphie, Pennsylvanie, États-Unis.

REYNAR J. C. (1994). An Automatic Method of Finding Topic Boundaries. *Dans Proc. of the 32nd annual meeting on Association for Computational Linguistics (ACL)*, Las Cruces, Nouveau Mexique.

REYNAR J. C. (1998). *Topic Segmentation: Algorithms and Applications.* Thèse de doctorat, University of Pennsylvania, Philadelphie, Pennsylvanie, États-Unis.

REYNAR J. C. (1999). Statistical Models for Topic Segmentation. *Dans Proc. of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, College Park, Maryland, États-Unis.

RICCARDI G., PIERACCINI R. et BOCCHIERI E. (1996). Stochastic Automata for Language Modeling. *Computer Speech and Language*, 10(4):265–293.

RICHMOND K., SMITH A. et AMITAY E. (1997). Detecting Subject Boundaries within Text: A Language Independent Statistical Approach. *Dans Proc. of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Providence, Rhode Island, États-Unis.

RIES K., BUØ F. D. et WAIBEL A. (1996). Class Phrase Models for Language Modeling. *Dans Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie, Pennsylvanie, États-Unis.

RISTAD E. S. (1995). A Natural Law of Succession. Rapport technique, Princeton University, New Jersey, États-Unis.

ROARK B. (2001). Probabilistic Top-Down Parsing and Language Modelling. *Computational Linguistics*, 27(2):249–276.

ROSENFELD R. (1996). A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language*, 10:187–228.

ROSENFELD R. (1997). A Whole Sentence Maximum Entropy Language Model. *Dans Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Santa Barbara, Californie, États-Unis.

ROSENFELD R. (2000a). Incorporating Linguistic Structure into Statistical Language Models. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 358:1311–1324.

ROSENFELD R. (2000b). Two Decades of Statistical Language Modeling: Where do we Go from Here? *Proc. of the IEEE*, 88(8):1270–1278.

ROSSIGNOL M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle.* Thèse de doctorat, Université de Rennes 1, Rennes, France.

ROSSIGNOL M. et SÉBILLOT P. (2003). Extraction statistique sur corpus de classes de mots-clés thématiques. *Traitement Automatique des Langues (TAL)*, 44(3):217–246.

ROSSIGNOL M. et SÉBILLOT P. (2006). Acquisition sur corpus non spécialisés de classes sémantiques thématisées. *Dans Actes des 8èmes journées internationales d'analyse statistique des données textuelles (JADT)*, Besançon, France.

RUEBER B. (1997). Obtaining Confidence Measures from Sentence Probabilities. *Dans Proc. of the 5th European Conference on Speech, Communication, Technology (Eurospeech)*, Rhodes, Grèce.

SAKO A., TAKIGUCHI T. et ARIKI Y. (2007). Language Modeling Using PLSA-Based Topic HMM. *Dans Proc. of the 10th European Conference on Speech Communication and Technology (Eurospeech)*, Anvers. Belgique.

SALTON G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co.

SALTON G., SINGHAL A., BUCKLEY C. et MITRA M. (1996). Automatic Text Decomposition Using Text Segments and Text Themes. *Dans Proc. of the the 7th ACM conference on Hypertext*, Bethesda, Maryland, États-Unis.

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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Phoenix, Arizona, États-Unis.

SAVOY J. (1999). A Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science*, 50(10):944–952.

SCHEYTT P., GEUTNER P. et WAIBEL A. (1998). Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain. *Dans Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, Seattle, Wahington, États-Unis.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Dans Proc. of the International Conference on New Methods in Language Processing*, Manchester, Royaume-Uni.

SCHMID H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Dans Proc. of the ACL SIGDAT Workshop*, Dublin, Irlande.

SCHUKAT-TALAMAZZINI E. G., HENDRYCH R., KOMPE R. et NIEMANN H. (1995). Permugram Language Models. *Dans Proc. of the 4th European Conference on Speech Communication and Technology (Eurospeech)*, volume 3, Madrid, Espagne.

SCHWARTZ R. et AUSTIN S. (1990). Efficient, High-Performance Algorithms for N-Best Search. *Dans Proc. of the DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvanie, États-Unis.

SCHWARTZ R. et AUSTIN S. (1991). A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Toronto, Canada.

SEBASTIANI F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

SENEFF S. (1992). TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics*, 18(1):61–86.

SENEFF S., MCCANDLESS M. et ZUE V. (1995). Integrating Natural Language into the Word Graph Search for Simultaneous Speech Recognition and Understanding. *Dans Proc. of the 4th European Conference on Speech Communication and Technology (Eurospeech)*, Madrid, Espagne.

SENEFF S., WANG C. et HAZEN T. J. (2003). Automatic Induction of N-Gram Language Models from a Natural Language Grammar. *Dans Proc. of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Genève, Suisse.

SETHY A., GEORGIOU P. G. et NARAYANAN S. (2005). Buildling Topic Specific Language Models from Webdata Using Competitive Models. *Dans Proc. of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, Lisbonne, Portugal.

SEYMORE K. et ROSENFELD R. (1997). Using Story Topics for Language Model Adaptation. *Dans Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Grèce.

SHANNON C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, 30:50–64.

SHRIBERG E. (2001). To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

SIU M.-H. et OSTENDORF M. (2000). Variable N-Grams and Extensions for Conversational Speech Language Modeling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63–75.

STOKES N., CARTHY J. et SMEATON A. F. (2004). SeLeCT: A Lexical Cohesion Based News Story Segmentation System. *AI Communications*, 17(1):3–12.

STOLCKE A. (1995). An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities. *Computational Linguistics*, 21(2):165–202.

STOLCKE A. (2002). SRILM - An Extensible Language Modeling Toolkit. *Dans Proc. of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, États-Unis.

Stolcke A., König Y. et Weintraub M. (1997). Explicit Word Error Minimization In N-Best List Rescoring. *Dans Proc. of the 5th European Conference on Speech, Communication, Technology (Eurospeech)*, volume 1, Rhodes, Grèce.

Stolcke A. et Segal J. (1994). Precise N-Gram Probabilities from Stochastic Context-Free Grammars. *Dans Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Las Cruces, Nouveau Mexique, États-Unis.

Stolcke A. et Shriberg E. (1996). Statistical Language Modeling for Speech Disfluencies. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Atlanta, Géorgie, États-Unis.

Stolcke A., Shriberg E., Bates R., Ostendorf M., Hakkani D., Plauche M., Tür G. et Lu Y. (1998). Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words. *Dans Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)*, volume 5, Sydney, Australie.

Su K.-Y., Chiang T.-H. et Lin Y.-C. (1992). A Unified Framework to Incorporate Speech and Language Information in Spoken Language Processing. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, San Francisco, Californie, États-Unis.

Suhm B. et Waibel A. (1994). Towards Better Language Models for Spontaneous Speech. *Dans Proc. of the 3rd International Conference on Spoken Language Processing (ICSLP)*, volume 2, Yokohama, Japon.

Suzuki M., Kajiura Y., Ito A. et Makino S. (2006). Unsupervised Language Model Adaptation Based on Automatic Text Collection from WWW. *Dans Proc. of Interspeech*, Pittsburgh, Pennsylvanie, États-Unis.

Tillmann C. et Ney H. (1997). Word Triggers and the EM Algorithm. *Dans Proc. of the Workshop Computational Natural Language Learning (CoNLL)*, Madrid, Espagne.

Tür G., Hakkani-Tür D., Stolcke A. et Shriberg E. (2001). Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*, 21(1):31–57.

Uchimoto K., Nobata C., Yamada A., Sekine S. et Isahara H. (2002). Morphological Analysis of the Spontaneous Speech Corpus. *Dans Proc. of the 19th International Conference on Computational Linguistics (COLING)*, volume 2, Taipei, Taiwan.

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. *Dans Proc. of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, Toulouse, France.

Valli A. et Véronis J. (1999). Étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133.

VAN EYNDE F., ZAVREL J. et DAELEMANS W. (2000). Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. *Dans Proc. of the Conference on Language Resources and Evaluation (LREC)*, Athènes, Grèce.

VAN MULBREGT P., CARP I., GILLICK L., LOWE S. et YAMRON J. (1999). Segmentation of Automatically Transcribed Broadcast News Text. *Dans Proc. of the DARPA Broadcast News Workshop*, Herndon, Virginie, États-Unis.

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)*, Keystone, Colorado, États-Unis.

VERGNE J. (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique non combinatoire. Synthèse et résultats.* Habilitation à diriger des recherches, Université de Caen, Caen, France.

VERGYRI D., KIRCHHOFF K., DUH K. et STOLCKE A. (2004). Morphology-Based Language Modeling for Arabic Speech Recognition. *Dans Proc. of the 8th International Conference on Spoken Language Processing (ICSLP)*, île de Jeju, Corée du Sud.

VÉRONIS J. (2000). *Ingénierie des langues*, chapitre Annotation automatique de corpus : panorama et état de la technique, pages 111–129. Paris : Éditions Hermès.

VÉRONIS J. (2004). Le traitement automatique des corpus oraux. *Traitement Automatique des Langues (TAL)*, 45(2):7–14.

VITERBI A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2): 260– 269.

WANG W. et HARPER M. P. (2002). The SuperARV Language Model: Investigating the Effectiveness of Tightly Integrating Multiple Knowledge Sources. *Dans Proc. of the Empirical Methods in Natural Language Processing Conference (EMNLP)*, Philadelphie, Pennsylvanie, États-Unis.

WANG W., HARPER M. P. et STOLCKE A. (2003a). The Robustness of an Almost-Parsing Language Model Given Errorful Training Data. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Hong Kong, Chine.

WANG W. et STOLCKE A. (2007). Integrating MAP, Marginals, and Unsupervised Language Model Adaptation. *Dans Proc. of the 10th European Conference on Speech Communication and Technology (Eurospeech)*, Anvers, Belgique.

Wang W., Stolcke A. et Harper M. P. (2004). The Use of a Linguistically Motivated Language Model in Conversational Speech Recognition. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Montréal, Canada.

Wang Y.-Y., Acero A. et Chelba C. (2003b). Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. *Dans Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, îles Vierges, États-Unis.

Wang Y.-Y., Mahajan M. et Huang X. (2000). A Unified Context-Free Grammar and N-Gram Model for Spoken Language Processing. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, Istanbul, Turquie.

Weintraub M., Aksu Y., Dharanipragada S., Khudanpur S., Ney H., Prange J., Stolcke A., Jelinek F. et Shriberg E. (1996). LM95 project report: Fast training and portability. Rapport technique, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, États-Unis.

Wessel F., Schlüter R., Macherey K. et Ney H. (2001). Confidence Measures for Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298.

Wessel F., Schlüter R. et Ney H. (2000). Using Posterior Word Probabilities for Improved Speech Recognition. *Dans Proc. of IEEE International Conference on the Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, Istanbul, Turquie.

Whittaker E. W. D. (2000). *Statistical Language Modelling for Automatic Speech Recognition of Russian and English.* Thèse de doctorat, University of Cambridge, Cambridge, Royaume-Uni.

Witten I. H. et Bell T. C. (1991). The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.

Witten I. H. et Frank E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2ème édition.

Woodland P. C. (2001). Speaker Adaptation for Continuous Density HMMs: A Review. *Dans Proc. of ITRW on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France.

Wu J. et Khudanpur S. (1999). Combining Nonlocal, Syntactic and N-Gram Dependencies in Language Modeling. *Dans Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hongrie.

Yaari Y. (1997). Segmentation of Expository Texts by Hierarchical Agglomerative Clustering. *Dans Proc. of Recent Advances in Natural Language Processing (RANLP)*, Tzigov Chark, Bulgarie.

Yvon F., Boula De Mareüil P. B., D'Allessandro C., Auberge V., Bagein M., Bailly G., Béchet F., Foukia S., Goldman J.-P., Keller É., O'Shaughnessy D., Pagel V., Sannier F., Véronis J. et Zellner B. (1998). Objective Evaluation of Grapheme to Phoneme Conversion for Text-to-Speech Synthesis in French. *Computer Speech and Language*, 12(4):393–410.

Zechner K. et Waibel A. (1998). Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition. *Dans Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, Montréal, Canada.

Zhu X. et Rosenfeld R. (2001). Improving Trigram Language Modeling with the World Wide Web. *Dans Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, Salt Lake City, Utah, États-Unis.

Zitouni I., Smaïli K. et Haton J.-P. (2003). Statistical Language Modeling Based on Variable-Length Sequences. *Computer Speech and Language*, 17(1):27–41.

Annexe A

Jeux d'étiquettes utilisés

Cette annexe présente les cinq jeux d'étiquettes morpho-syntaxiques que nous avons employés au cours de nos expérimentations sur l'apport des POS pour améliorer la transcription automatique. Chaque ligne donne le code de l'étiquette associée à sa signification.

A.1 Étiquettes de base *norm*

Ce jeu comporte 93 étiquettes.

A.1.1 Adjectifs qualificatifs

ADJMS (masculin singulier)
ADJFS (féminin singulier)
ADJMP (masculin pluriel)
ADJFP (féminin pluriel)

A.1.2 Adverbes

ADV

A.1.3 Cardinaux (adjectifs numériques et déterminants)

CAR

A.1.4 Conjonctions

COO (conjonctions de coordination)
SUB (conjonctions de subordination)

A.1.5 Déterminants

DETMS (masculin singulier)
DETFS (féminin singulier)
DETMP (masculin pluriel)
DETFP (féminin pluriel)
DETPIG (pluriel indifférencié sur le genre)

A.1.6 Interjections

INT

A.1.7 Noms communs

NCMS (masculin singulier)
NCFS (féminin singulier)
NCMP (masculin pluriel)
NCFP (féminin pluriel)

A.1.8 Noms propres

NPMS (masculin singulier)
NPFS (féminin singulier)
NPMP (masculin pluriel)
NPFP (féminin pluriel)
NPSIG (singulier indifférencié sur le genre)
NPPIG (pluriel indifférencié sur le genre)
NPI (indifférencié sur le genre et le nombre)

A.1.9 Prépositions

PREP

A.1.10 Pronoms

Pronoms personnels

PPER1S (1ère personne du singulier)
PPER2S (2ème personne du singulier)
PPER3S (3ème personne du singulier)
PPER1P (1ère personne du pluriel)
PPER2P (2ème personne du pluriel)
PPER3P (3ème personne du pluriel)

Pronoms relatifs

PRMS (masculin singulier)
PRFS (féminin singulier)
PRMP (masculin pluriel)
PRFP (féminin pluriel)
PRI (indifférencié sur le genre et le nombre)

Pronoms indéfinis, démonstratifs, possessifs, cardinaux et interrogatifs

PMS (masculin singulier)
PFS (féminin singulier)
PMP (masculin pluriel)
PFP (féminin pluriel)

A.1.11 Verbes

Infinitif

VINF

Indicatif présent

VINDP1S (1ère personne du singulier)
VINDP2S (2ème personne du singulier)
VINDP3S (3ème personne du singulier)
VINDP1P (1ère personne du pluriel)
VINDP2P (2ème personne du pluriel)
VINDP3P (3ème personne du pluriel)

Indicatif imparfait

VINDI1S (1ère personne du singulier)
VINDI2S (2ème personne du singulier)
VINDI3S (3ème personne du singulier)
VINDI1P (1ère personne du pluriel)
VINDI2P (2ème personne du pluriel)
VINDI3P (3ème personne du pluriel)

Indicatif passé simple

VINDPS1S (1ère personne du singulier)
VINDPS2S (2ème personne du singulier)
VINDPS3S (3ème personne du singulier)
VINDPS1P (1ère personne du pluriel)
VINDPS2P (2ème personne du pluriel)
VINDPS3P (3ème personne du pluriel)

Indicatif futur simple

VINDF1S (1ère personne du singulier)
VINDF2S (2ème personne du singulier)
VINDF3S (3ème personne du singulier)
VINDF1P (1ère personne du pluriel)
VINDF2P (2ème personne du pluriel)
VINDF3P (3ème personne du pluriel)

Subjonctif présent

VSUBP1S (1ère personne du singulier)
VSUBP2S (2ème personne du singulier)
VSUBP3S (3ème personne du singulier)
VSUBP1P (1ère personne du pluriel)
VSUBP2P (2ème personne du pluriel)
VSUBP3P (3ème personne du pluriel)

Subjonctif imparfait

VSUBI1S (1ère personne du singulier)
VSUBI2S (2ème personne du singulier)
VSUBI3S (3ème personne du singulier)
VSUBI1P (1ère personne du pluriel)
VSUBI2P (2ème personne du pluriel)
VSUBI3P (3ème personne du pluriel)

Conditionnel présent

VCONP1S (1ère personne du singulier)
VCONP2S (2ème personne du singulier)
VCONP3S (3ème personne du singulier)
VCONP1P (1ère personne du pluriel)
VCONP2P (2ème personne du pluriel)
VCONP3P (3ème personne du pluriel)

Impératif présent

VIMPP2S (2ème personne du singulier)
VIMPP1P (1ère personne du pluriel)
VIMPP2P (2ème personne du pluriel)

Participe présent

VPARPRES

Participe passé

VPARPMS (masculin singulier)
VPARPFS (féminin singulier)
VPARPMP (masculin pluriel)
VPARPFP (féminin pluriel)

A.1.12 Symbole

SYMBOLE

Cette classe regroupe des mots représentant des caractères, tels que « *arobase* », « *tiret* » ou « *point* ». Elle est notamment utilisée dans le contexte des adresses de courriels ou de sites Web.

A.2 Premier ensemble réduit d'étiquettes *red*1

Ce jeu comporte 30 étiquettes. Il effectue peu de distinctions entre les POS afin d'évaluer si les informations morphologiques sont prépondérantes pour discriminer les hypothèses contenant des erreurs.

A.2.1 Adverbes

ADV

A.2.2 Cardinaux (adjectifs numériques et déterminants)

CAR

A.2.3 Conjonctions

COO (conjonctions de coordination)
SUB (conjonctions de subordination)

A.2.4 Déterminants

DETMS (masculin singulier)
DETFS (féminin singulier)
DETMP (masculin pluriel)
DETFP (féminin pluriel)
DETPIG (pluriel indifférencié sur le genre)

A.2.5 Interjections

INT

A.2.6 Prépositions

PREP

A.2.7 Adjectifs qualificatifs, noms communs, noms propres et pronoms (indéfinis, démonstratifs, possessifs, cardinaux et interrogatifs)

Ils ne sont plus distingués que par le genre et le nombre :
MS (masculin singulier)
FS (féminin singulier)
MP (masculin pluriel)
FP (féminin pluriel)
SIG (singulier indifférencié sur le genre)
PIG (pluriel indifférencié sur le genre)
I (indifférencié sur le genre et en nombre)

A.2.8 Pronoms relatifs

Nous avons choisi d'associer des classes particulières à ces pronoms dans la mesure où ils jouent un rôle structurant dans une phrase. PRMS (masculin singulier)
PRFS (féminin singulier)
PRMP (masculin pluriel)
PRFP (féminin pluriel)
PRI (indifférencié sur le genre et le nombre)

A.2.9 Pronoms personnels et verbes

Ils ne sont plus distingués que par la personne et le genre :
1S (1ère personne du singulier)
2S (2ème personne du singulier)
3S (3ème personne du singulier)
1P (1ère personne du pluriel)
2P (2ème personne du pluriel)
3P (3ème personne du pluriel)

A.2.10 Symbole

SYMBOLE

A.3 Deuxième ensemble réduit d'étiquettes *red*2

Ce jeu comporte 70 étiquettes. À l'exception des verbes, toutes les étiquettes utilisées dans *norm* demeurent inchangées.

A.3.1 Auxiliaire « *avoir* » et « *être* »

Des étiquettes spécifiques sont créées pour les verbes « *avoir* » et « *être* ». Elles sont attribuées indifféremment selon qu'ils soient employés en tant qu'auxiliaire ou en tant que verbe principal.

Infinitif

AVOIRINF
ETREINF

Participe présent

AVOIRPARPRES
ETREPARPRES

Participe passé

VPARPMS (masculin singulier)
VPARPFS (féminin singulier)
VPARPMP (masculin pluriel)
VPARPFP (féminin pluriel)

Formes conjuguées

AVOIR1S (1ère personne du singulier)
AVOIR2S (2ème personne du singulier)
AVOIR3S (3ème personne du singulier)
AVOIR1P (1ère personne du pluriel)
AVOIR2P (2ème personne du pluriel)
AVOIR3P (3ème personne du pluriel)
ETRE1S (1ère personne du singulier)
ETRE2S (2ème personne du singulier)
ETRE3S (3ème personne du singulier)
ETRE1P (1ère personne du pluriel)
ETRE2P (2ème personne du pluriel)
ETRE3P (3ème personne du pluriel)

A.3.2 Autres verbes

Infinitif

VINF

Participe présent

VPARPRES

Participe passé

VPARPMS (masculin singulier)
VPARPFS (féminin singulier)
VPARPMP (masculin pluriel)
VPARPFP (féminin pluriel)

Formes conjuguées

V1S (1ère personne du singulier)
V2S (2ème personne du singulier)
V3S (3ème personne du singulier)
V1P (1ère personne du pluriel)
V2P (2ème personne du pluriel)
V3P (3ème personne du pluriel)

A.4 Premier ensemble étendu d'étiquettes *ext*1

Ce jeu comporte 167 étiquettes. Il conserve toutes les étiquettes de *norm* en les étendant par des étiquettes spécifiques pour les mots les plus courants d'après notre corpus d'apprentissage : « *à* », « *après* », « *au* », « *autre* », « *autres* », « *aux* », « *avant* », « *avec* », « *c'* », « *car* », « *ce* », « *certains* », « *chez* », « *cinq* », « *comme* », « *comment* », « *contre* », « *d'* », « *de* », « *d'autres* », « *dans* », « *depuis* », « *des* », « *deux* », « *devant* », « *donc* », « *dont* », « *du* », « *en* », « *entre* », « *et* », « *jusqu'* », « *l'* », « *la* », « *le* », « *les* », « *leur* », « *leurs* », « *mais* », « *même* », « *ou* », « *où* », « *par* », « *parce qu'* », « *parce que* », « *pendant* », « *plusieurs* », « *pour* », « *première* », « *près* », « *puis* », « *puisque* », « *qu'* », « *quand* », « *quatre* », « *que* », « *qui* », « *quoi* », « *s'* », « *sans* », « *se* », « *selon* », « *si* », « *soit* », « *son* », « *sous* », « *sur* », « *tous* », « *tout* », « *toute* », « *toutes* », « *trois* », « *un* », « *une* », « *vers* », « *y* ». Les mots les plus courants considérés comme étiquettes ont été déterminés en gardant parmi les 100 plus fréquents ceux qui n'étaient pas des pronoms personnels ambigus.

N.B. : Une même classe a été associée aux mots possédant la même ambiguïté au niveau des classes grammaticales et différant uniquement par l'élision, *i.e.*, « *d'* » et « *de* », « *qu'* » et « *que* », « *parce qu'* » et « *parce que* ».

A.5 Deuxième ensemble étendu d'étiquettes *ext*2

Ce jeu comporte 144 étiquettes. Il conserve toutes les étiquettes de *red*2 en les étendant par des étiquettes spécifiques pour les mots grammaticaux les plus courants de notre corpus d'apprentissage. La liste de ces mots est la même que celle précédemment présentée pour *ext*1.

Annexe B

Paramétrage de l'étiqueteur morpho-syntaxique

Les paramètres de notre étiqueteur morpho-syntaxique, à savoir le type de lissage des probabilités lexicales et contextuelles et l'ordre N du contexte pris en compte dans le calcul de $P(w_i|w_{i-N+1}^{i-1})$, ont été fixés sur un corpus de 40 minutes — soit 6 780 mots — transcrit manuellement.

B.1 Étiquettes de base *norm*

Les tableaux B.1, B.2 et B.3 reportent les pourcentages d'étiquettes correctes obtenus en faisant varier les paramètres du lissage des probabilités calculées par l'étiqueteur.

Une comparaison des valeurs obtenues montre que les paramètres influent peu sur la qualité d'étiquetage. On remarque toutefois que le nombre d'étiquettes correctes est globalement plus élevé en utilisant un lissage additif avec $\delta = 1$ pour les probabilités lexicales, et un lissage de Katz avec un ordre de 3 pour les probabilités contextuelles. Nous employons donc ces paramètres de lissage pour construire notre étiqueteur.

Ordre	1	2	3	4	5	6	7	8
Pas de lissage	90,49	93,81	94,03	94,00	93,94	93,94	93,99	94,00
Katz	90,49	93,84	**94,12**	94,06	94,03	93,97	93,96	93,93
Witten-Bell	90,49	93,86	93,99	94,03	93,90	93,94	93,93	93,93
Kneser-Ney	90,49	93,81	93,99	93,88	93,94	93,93	94,00	93,99
Kneser-Ney modifié	90,49	93,80	94,05	93,94	93,71	93,72	93,66	93,60
Ristad	90,49	93,84	94,03	94,05	94,00	93,94	94,02	94,00

Tab. B.1: *Pourcentage d'étiquettes correctes avec un lissage de Good-Turing pour les probabilités lexicales*

Ordre	1	2	3	4	5	6	7	8
Pas de lissage	90,13	93,91	94,12	93,93	93,93	93,96	93,97	93,97
Katz	90,13	93,93	**94,19**	94,00	93,96	93,99	93,87	93,83
Witten-Bell	90,13	93,94	94,11	94,00	93,93	93,93	93,93	93,93
Kneser-Ney	90,13	93,91	94,08	93,93	93,90	93,93	94,02	93,97
Kneser-Ney modifié	90,13	93,91	94,05	93,88	93,71	93,74	93,63	93,55
Ristad	90,13	93,93	94,12	93,99	93,94	93,90	93,96	93,97

TAB. B.2: *Pourcentage d'étiquettes correctes avec un lissage additif ($\delta = 0,5$) pour les probabilités lexicales*

Ordre	1	2	3	4	5	6	7	8
Pas de lissage	90,01	93,97	94,14	94,00	93,94	93,97	93,97	93,97
Katz	90,01	93,97	**94,22**	94,03	93,94	93,93	93,83	93,78
Witten-Bell	90,01	93,97	94,17	94,03	93,91	93,88	93,87	93,86
Kneser-Ney	90,01	93,94	94,12	94,03	93,80	93,99	94,05	94,03
Kneser-Ney modifié	90,01	93,91	94,09	94,05	93,69	93,71	93,57	93,50
Ristad	90,01	93,99	94,17	94,09	93,99	93,96	93,94	93,93

TAB. B.3: *Pourcentage d'étiquettes correctes avec un lissage additif ($\delta = 1$) pour les probabilités lexicales*

B.2 Autres ensembles d'étiquettes *red*1, *red*2, *ext*1 et *ext*2

En réitérant les mêmes expériences sur les quatre autres jeux d'étiquettes pour tester l'influence des paramètres, nous avons obtenu des résultats similaires à ceux obtenus précédemment. Néanmoins, pour l'ensemble *red*1 qui contient le moins d'étiquettes, nous avons obtenu de meilleurs résultats en utilisant un ML quadrigrammes pour les probabilités contextuelles. Nous employons donc un lissage additif ($\delta = 1$) pour les probabilités lexicales et un lissage de Katz pour les probabilités contextuelles lors de l'étiquetage par *red*1, *red*2, *ext*1 et *ext*2. De plus, nous tenons compte d'un contexte de 4 étiquettes en ce qui concerne l'ensemble *red*1, et d'un contexte de 3 pour *red*2, *ext*1 et *ext*2.

Annexe C

Exemples de cooccurrents

Nous donnons dans cette annexe une liste des paires de lemmes qui obtiennent des valeurs élevées d'information mutuelle sur le corpus *Le Monde* limité aux années 1999 à 2003. Seules les classes utilisées lors de la segmentation thématique (les noms communs, les noms propres, les adjectifs et les verbes) et les couples apparaissant ensemble au moins 15 fois dans des fenêtres d'analyse de 20 mots ont été conservés.

Au vu des 25 lemmes les plus proches selon le réseau de cooccurrences pour quatre lemmes différents étudiés (Tab. C.1 et C.2), il apparaît que les liens les plus forts sont souvent d'ordre syntagmatique, *e.g.* « *radar* » et « *anti-aérien* » ou « *élection* » et « *proportionnel* ». On note aussi une surreprésentation des noms propres ou des sigles, la plupart ayant peu d'occurrences et apparaissant dans des contextes très précis.

irak			radar		
mot	occ	coh	mot	occ	coh
après-saddam	75	0,198	awac	29	0,275
vx	35	0,196	intercepteur	17	0,261
unscom	116	0,194	anti-aérien	17	0,252
bremer	121	0,193	sol-air	21	0,233
adm	69	0,190	antiaérien	26	0,229
désarmement	710	0,189	infrarouge	24	0,220
bassorah	28	0,189	alaska	21	0,219
américano-britannique	127	0,188	furtif	22	0,215
blix	243	0,186	détection	33	0,208
koweït	434	0,184	missile	138	0,203
garner	48	0,184	laser	20	0,199
chaldéen	16	0,183	antimissile	26	0,198
endiguement	16	0,180	automatique	57	0,194
scud	38	0,178	détecter	37	0,194
libyer	21	0,177	adverse	17	0,194
kirkouk	60	0,176	satellite	108	0,194
ritter	18	0,176	batterie	35	0,192
anglo-américain	92	0,176	avion	161	0,183
hakim	62	0,175	optique	27	0,182
saddam	1347	0,174	embarquer	27	0,180
kurdistan	189	0,174	raid	27	0,180
hussein	1188	0,174	alerte	39	0,180
mossoul	59	0,174	météo	19	0,179
iran	915	0,173	aviation	24	0,178
barzani	24	0,173	engin	24	0,177

TAB. C.1: *25 lemmes les plus proches d'après le réseau de cooccurrences pour deux lemmes « irak » et « radar » apparaissant dans l'une des émissions d'information à segmenter. occ représente le nombre d'occurrences de chaque couple de lemmes dans une fenêtre d'analyse et coh leur information mutuelle normalisée*

syndicat			scrutin		
mot	occ	coh	mot	occ	coh
omnipraticien	34	0,201	uninominal	80	0,244
cisl	63	0,201	proportionnel	388	0,208
snj	172	0,199	bipolarisation	20	0,200
snalc	46	0,198	sénatorial	103	0,185
metall	159	0,198	boycotter	63	0,182
dag	20	0,197	cantonal	88	0,182
fsir	125	0,197	invalider	33	0,180
geneix	29	0,196	rdr	32	0,180
snep	100	0,196	irrégularité	74	0,176
eyrolle	60	0,194	entacher	40	0,175
unser	208	0,193	toledo	17	0,174
sna	18	0,192	majoritaire	270	0,173
fdsea	60	0,190	alassane	21	0,172
sne	286	0,188	dépouillement	38	0,172
fen	130	0,187	pourvoir	33	0,171
spi	36	0,186	frauduleux	41	0,170
tuc	27	0,186	sincérité	57	0,170
sylviculteur	19	0,186	législatif	463	0,170
voirin	20	0,186	abstentionniste	20	0,169
ufap	19	0,185	urne	94	0,168
baro	26	0,185	électeur	438	0,168
sud-ptt	66	0,185	abstention	140	0,167
solidarnosc	26	0,185	timorais	17	0,166
sgen	111	0,184	votant	36	0,166
unsa	223	0,184	mode	770	0,165

TAB. C.2: *25 lemmes les plus proches d'après le réseau de cooccurrences pour deux lemmes « syndicat » et « scrutin » apparaissant dans l'une des émissions d'information à segmenter. occ et coh gardent la même signification que la table C.1*

Annexe D

Exemple de segmentation d'une émission d'information

Cette annexe fournit un exemple de segmentation thématique d'un extrait d'un quart d'heure d'une émission d'information diffusée sur France Inter et faisant partie du corpus de test. Nous y présentons la segmentation de référence obtenue à partir d'une transcription manuelle (à gauche sur la figure donnée ci-après), et une segmentation automatique effectuée sur une sortie du système de RAP (à droite sur la figure). Ce dernier découpage a été obtenu en prenant en compte le modèle de cohésion lexicale présenté à la section 7.3, des indices acoustiques (*cf.* section 7.6.1) et un modèle 6-grammes cachés basé sur les POS (*cf.* section 7.6.2), en recourant au modèle multi-sources que nous proposons en section 7.6.

Les deux versions de segmentation présentées ici sont alignées au niveau du groupe de souffle, dont la numérotation est donnée en début de chaque transcription. Les régions non grisées correspondent à des *fillers* qui sont ignorés lors de l'évaluation de la qualité de la segmentation. Les parties inscrites en italique sont quant à elles des annonces de titre de l'émission. Les barres horizontales de la segmentation de référence matérialisent les points de rupture thématique ; ceux étant optionnels car correspondant aux passages d'un sous-thème à un autre (par exemple entre deux sujets relatifs à la guerre en Irak) sont représentés par des traits fins. Les lignes horizontales relatives à la transcription automatique indiquent les frontières proposées par le modèle multi-sources.

0 france inter il est huit heures	*0* france-inter il est huit heures
1 le journal patrick roger	*1* le journal patrick roger
2 bonjour	*2* bonjour
3 l'heure de la retraite a vraiment sonné rencontres capitales aujourd'hui et mardi	*3 l' heure de la retraite a vraiment sonné rencontre capitale aujourd'hui et mardi*
4 les ministres françois fillon et jean-paul delevoye dévoilent leurs batteries aux partenaires sociaux	*4 les ministres françois fillon et jean-paul delevoye dévoilent leurs batteries aux partenaires sociaux*
5 mécontents des mesures pour les fonctionnaires les syndicats guettent maintenant les arbitrages pour le secteur privé	*5 mécontent des mesures pour les fonctionnaires les syndicats qui est maintenant les arbitrages pour le secteur privé*

6 un an après que reste-t-il du vingt et un avril journée spéciale sur france inter	*6 un an après que reste -t-il du vingt-et-un avril journée spéciale sur france-inter*
7 trois invités d'ici neuf heures jean-marie le pen françois hollande et nicolas sarkozy avant cela	*7 trois invités dix six neuf heures jean-marie le pen françois hollande et nicolas sarkozy avant cela*
8 un oeil dans le rétro avec les vrais chiffres de ce premier tour de la présidentielle	*8 un oeil dans le rétro avec les vrais chiffres de ce premier tour de la présidentielle*
9 le détail et l'analyse de jean-michel apathie	*9 le détail des analystes jean-michel aphatie*
10 une semaine après la vraie fausse fin de la guerre que reste-t-il à trouver pour les américains en irak	*10 une semaine après la vraie aux fin de la guerre que reste -t-il à trouver pour les américains en irak*
11 des armes de destruction massive toujours aucune trace mais un millier de spécialistes envoyés sur le terrain par washington	*11 des armes de destruction massive toujours aucune trace des milliers de spécialistes envoyés sur le terrain par washington*
12 où est saddam hussein avec son entourage peut-être une piste avec euh l'arrestation	*12 où est saddam hussein ainsi que son entourage peut être une piste avec l' arrestation*
13 du demi-frère du président irakien hier nous irons également à bagdad	*13 du demi-frère du président irakien et hier nous irons également à bagdad*
14 où la rue est divisée sur l'attitude à adopter face aux militaires américains qui encadrent la vie quotidienne	*14 où l' armée est divisée sur l' attitude à adopter face aux militaires américains qui encadrent la vie quotidienne*
15 il aurait pu en tirer une oeuvre une vente hors norme que la vente aux enchères de la collection d'andré breton	*15 s' il aurait pu retirer une oeuvre uniquement hors norme que la vente aux enchères de la collection d' andré breton*
16 qui s'est achevée hier elle a rapporté beaucoup plus d'argent que prévu	*16 s' est achevé hier est emporté beaucoup plus argent*
17 souriez vous conduisez et vous êtes filmés	*17 sauriez vous conduisaient que vous êtes filmés*
18 journée de la courtoisie au volant et déploiement de nouveaux radars sur les routes de france	*18 journée de la courtoisie au volant et déploiement de nouveaux radars sur les routes de france*
19 les caméras numériques sur le bord des routes enfin nous terminerons sur un air de vacances	*19 les caméras numériques sur le bord des axes de francs fin terminons sur un air de vacances*
20 les pieds dans l'eau un peu fraîche je vous l'accorde à boulogne-sur-mer	*20 les pieds dans le lointain fait je vous l' accorde à boulogne charme*
21 france inter	*21* france-inter
22 prions pour que ça passe	*22* prions pour que ça passe
23 c'est en ce vendredi saint ce que doivent se dire les ministres françois fillon et jean-paul delevoye	*23* et en ce vendredi cinq ce que doit se dire les ministres françois fillon et jean-paul delevoye
24 c'est la journée qu'ils ont choisi pour dévoiler la nouvelle bible des retraites en france	*24* c' est la journée qu' ils ont choisi pour dévoiler la nouvelle bible des retraites en france
25 rencontre déterminante donc aujourd'hui et mardi également au ministère du travail	*25* rencontre déterminante donc aujourd'hui les martyrs d' allemands au ministère du travail
26 avec la procession de tous les partenaires sociaux	*26* avec la procession de tous les partenaires sociaux
27 les syndicats et les organisations patronales vont être reçus chacun leur tour	*27* les syndicats et les organisations patronales vont être reçus chacun leur tour
28 la réforme des retraites est le dossier chaud du gouvernement ça passe ou ça casse	*28* la réforme des retraites et le dossier chaud du gouvernement ça passe ou ça casse
29 c'est l'épreuve de cette fin de cette fin de printemps et de début d'été	*29* c' est l' épreuve de ce fin de de cette fin de printemps et de début d' été
30 où le texte doit être voté à l'assemblée	*30* où le texte doit être voté à l' assemblée
31 alors que des informations ont été données au compte gouttes ces dernières semaines	*31* alors que des informations étaient données au compte-gouttes ces dernières semaines

32 mercredi le ministre de la fonction publique jean-paul delevoye a pris tout le monde de court	*32* mercredi le ministre de la fonction publique jean-paul delevoye a pris tout le monde de court
33 en donnant des détails sur ce qui va changer pour les fonctionnaires	*33* en donnant des détails sur ce qui va changer pour les fonctionnaires
34 aujourd'hui le secteur privé va être fixé voyons avec euh véronique julia	*34* aujourd'hui le secteur privé va être fixé voyons avec euh véronique jullien
35 ce que l'on sait pour l'instant des grandes lignes de la réforme	*35* ce que l' on sait pour l' instant des grandes lignes de la réforme
36 le gouvernement ne s'en cache pas pour financer les retraites il faudra faire des sacrifices	*36* le gouvernement ne s' engage pas pour financer les retraites il faudra faire des sacrifices
37 selon le principe d'équité les fonctionnaires vont devoir cotiser plus longtemps	*37* selon principes d' équité les fonctionnaires vont devoir cotiser plus longtemps
38 en deux mille huit ce sera donc quarante ans pour tous et même sans doute quarante deux ans d'ici deux mille vingt	*38* en deux mille huit ce sera donc quarante ans pour tous et même sans doute quarante-deux ans d' ici deux mille vingt
39 pour les salariés du privé pas question de revenir sur la réforme balladur	*39* pour les salariés du privé pas question de revenir sur la réforme balladur
40 d'ici cinq ans leurs retraites seront calculées sur la base de leurs vingt cinq meilleures années	*40* d' ici cinq ans leur retraite seront calculées sur la base de leur vingt-cinq meilleures années
41 qui conduira immanquablement à une baisse du niveau de leurs pensions	*41* ce qui conduira immanquablement à une baisse du niveau de leur pension
42 les plus bas salaires pourront-ils partir en retraite avec cent pour cent du smic	*42* les plus bas salaires pourront ils partir en retraite avec cent pour cent du smic
43 comme le réclament les syndicats pour l'instant mystère	*43* comme le réclament les syndicats pour l' instant mystère
44 on ne sait pas davantage si les salariés qui ont commencé à travailler très jeunes	*44* on ne sait pas davantage si les salariés qui ont commencé à travailler très jeunes
45 pourront bénéficier d'une retraite à taux plein en cas de départ avant soixante ans	*45* pourront bénéficier d' une retraite à taux plein en cas de départ avant soixante ans
46 en ce qui concerne les fonctionnaires le gouvernement mise surtout sur la durée d'activité	*46* en ce qui concerne les fonctionnaires le gouvernement mise surtout sur la durée d' activité
47 pour limiter les départs précoces une décote s'appliquera dès l'an prochain sur chaque année non cotisée	*47* pour limiter les départs précoces une décote s' appliquera dès l' an prochain sur chaque année n' ont cotisé
48 à l'inverse ceux qui partiront plus tard bénéficieront d'un bonus valable également pour les salariés du privé	*48* à l' inverse qui partiront plus tard bénéficieront d' un bonus valable également pour les salariés du privé
49 enfin pour compenser les efforts demandés aux fonctionnaires on ne touchera pas à leurs avantages familiaux	*49* enfin pour compenser les efforts demandés aux fonctionnaires qu' on ne touchera pas à leurs avantages familiaux
50 et les primes jusqu'à présent laissées de côté devraient être intégrées dans le calcul des pensions	*50* et les primes jusqu' à présent les sept côté devraient être intégrés dans le calcul des pensions
51 oui en tout cas les mesures ne font pas l'unanimité chez les syndicats	*51* mais on est en tout cas les mesures ne font pas l' unanimité chez les syndicats
52 qui attendent tout de même la fin des rendez-vous chez françois fillon et jean-paul delevoye	*52* qui attend tout de même la fin des rendez vous chez françois fillon et jean-paul delevoye
53 ils guettent d'éventuelles concessions avant de se prononcer plus précisément sur leurs actions	*53* ils guettent d' éventuelles concessions avant de se prononcer plus précisément sur leurs actions
54 ils se retrouveront mercredi pour une réunion intersyndicale et en fonction des mesures ils décideront ou non	*54* ils se retrouveront mercredi pour une réunion intersyndicale et en fonction des mesures ils décideront ou non

55 de journées de mobilisation dans les semaines à venir et de la tournure que pourrait prendre le défilé du premier mai	*55* deux journées de mobilisation dans les semaines à venir et de la tournure que pourrait prendre le défilé du premier mai
56 et pendant que ça menace de bouillonner dans la rue la classe politique se souvient du bouillon de l'année dernière	*56* et pendant que sa menace de bouillonne et dans la rue la classe politique se souvient du bouillon de l' année dernière
57 du séisme du vingt et un avril le premier tour de la présidentielle un an après	*57* du séisme du vingt-et-un avril le premier tour de la présidentielle un an après
58 toute la journée france inter s'interroge que reste-t-il de ce vingt et un avril	*58* toute la journée à france-inter s' interroge que reste -t-il de ce vingt-et-un avril
59 une date qui a donné le vertige à une partie de la france	*59* une date qui a donné le vertige à une partie de la france
60 avec des résultats présentés sans doute un peu trop vite et à tort jean-michel apathie	*60* avec des résultats présentés sans doute un peu trop vite y a tort jean-michel aphatie
61 comme le succès du front national face aux partis traditionnels	*61* comme le succès du front national face aux partis traditionnels
62 contrairement à l'idée reçue le vingt et un avril deux mille deux n'est pas un jour de triomphe pour le front national	*62* contrairement aux idées reçues le vingt-et-un avril deux mille deux n' est pas un jour de triomphe pour le front national
63 son candidat alors totalise quatre millions huit cents mille suffrages	*63* son candidat alors totalisent quatre millions huit cents mille suffrages
64 une progression infime par rapport au premier tour de la même élection en mille neuf cent quatre vingt quinze	*64* une progression infime par rapport au premier tour de la même élection en mille neuf cent quatre-vingt-quinze
65 où jean-marie le pen recueillait quatre millions six cents mille voix deux cents mille voix gagnées en sept ans ce n'est pas terrible	*65* jean-marie le pen a recueilli quatre millions six cents mille voix deux cent mille voix gagnant sept ans ce n' est pas terrible
66 en revanche dans le même laps de temps on constate un écroulement des socialistes et de la droite	*66* en revanche dans le même laps de temps on constate un écroulement des socialistes et de la droite
67 candidat en mille neuf cent quatre vingt quinze lionel jospin regroupait sur son nom sept millions de voix	*67* candidat en mille neuf cent quatre-vingt-quinze lionel jospin regrouper sur son nom sept millions de voix
68 il se contente de quatre millions six cents mille suffrages le vingt et un avril deux mille deux deux millions et demi d'électeurs perdus en sept ans	*68* il se contente de quatre millions six cents mille suffrages le vingt-et-un avril deux mille deux deux millions et demi d' électeurs perdus en sept ans
69 la droite républicaine elle totalisait douze millions de voix en mille neuf cent quatre vingt quinze	*69* la droite républicaine elle a totalisé deux millions d' euros en mille neuf cent quatre-vingt-quinze
70 addition des électorats chirac balladur elle se contente de huit millions d'électeurs en deux mille deux	*70* addition des électorats chirac balladur se contentent de huit millions d' électeurs en deux mille deux
71 addition des électorats chirac bayrou madelin quatre millions d'électeurs envolés en sept ans c'est un record	*71* l' addition des électorats chirac bayrou madelin quatre millions d' électeurs en voyant ces temps c' est un record
72 ajoutons à ce chiffre celui des abstentionnistes onze millions et demi de français	*72* ajoutons à ces chiffres celui des abstentionnistes en deux millions et demi de français
73 et l'on obtient la vérité du vingt et un avril deux mille deux	*73* mais en aucun la vérité du vingt-et-un avril deux mille deux
74 l'expression d'une méfiance sinon d'un mépris à l'égard d'un système politique malade et usé	*74* l' expression d' une méfiance sinon de mépris à l' égard d' un système politique malade et usé

75 producteur de davantage de méfiance que de confiance	*75* producteur de davantage de méfiance que de confiance
76 c'est cela qu'il faudrait changer mais un an après rien n'a changé	*76* c' est cela qu' il faudrait changer mais un an après rien
77 alors si rien n'a vraiment changé que faut-il faire euh stéphane paoli l'évoquera tout à l'heure	*77* alors si rien n' a vraiment changé que faut il faire euh stéphane paoli il évoquera tout à l' heure
78 à l'occasion d'une matinée spéciale trois invités jusqu'à neuf heures	*78* à l' occasion d' une matinée spéciale trois invités jusqu' à neuf heures
79 à huit heures vingt jean-marie le pen à huit heures quarante françois hollande et à huit heures cinquante nicolas sarkozy	*79* à huit heures vingt jean-marie le pen à huit heures quarante france hollande et à huit heures cinquante euh nicolas sarkozy
80 le téléphone sonne d'alain bédouet ce soir se penchera également sur les conséquences et les enseignements de ce vingt et un avril	*80* le téléphone sonne d' alain bédouet ce soir se penchera également sur les conséquences et les enseignements de ce vingt-et-un avril
81 vos questions au zéro un quarante cinq vingt quatre sept mille à partir de dix sept heures ou sur internet france-inter point com il y a d'ailleurs	*81* vos questions au zéro un quarante cinq vingt quatre sept mille à partir de dix-sept heures ou sur internet de france-inter pas comment il y a d' ailleurs
82 un forum spécifique sur ce thème encore un mot justement de la politique avec une côte d'azur	*82* un forum spécifique sur ce thème encore un mot justement de la politique avec une côte-d'azur
83 très prisée en ce week-end pascal le front national le ps et les verts vont se croiser	*83* très prisés en ce week-end pascal le front national le ps et les verts vont se croiser
84 à la veille du congrès du fn les deux partis de gauche organisent ce soir au même endroit à nice un meeting intitulé	*84* à la veille du congrès du fn les deux partis de gauche organise ce soir au même endroit indice un meeting intitulé
85 soirée de la démocratie	*85* soirée de la démocratie
86 il est huit heures six france inter c'est le journal de patrick roger	*86* il est huit heures six france-inter c' est le journal de patrick roger
87 sur la piste de saddam hussein en irak les militaires américains ne prononcent toujours pas le nom du président irakien	*87* sur la piste de saddam hussein en irak les militaires américains ne prononce toujours pas le nom du président irak
88 officiellement il n'est pas leur objectif le but c'est tout l'appareil de l'ancienne dictature et hier	*88* officiellement il n' est pas leur objectif le but c' est tout l' appareil de l' ancienne dictature est hier
89 ils ont réalisé une arrestation qui pourrait peut-être les conduire tout de même au président irakien puisqu'il s'agit	*89* ils ont réalisé une arrestation qui pourrait peut-être les conduire tout de même au président irakien puisqu' il s' agit
90 du demi-frère de saddam hussein barzan al-tikriti	*90* du demi-frère de saddam hussein barzane al tikrit
91 qui ces dernières années avait amassé une fortune colossale mais surtout illégale à genève jacques prévost	*91* qui ces dernières années avait amassé une fortune colossale mais surtout illégal à genève jacques prévost
92 c'est une gigantesque fortune accumulée depuis des années	*92* c' est une gigantesque fortune accumulée depuis des années
93 on dit ici que les avoirs du raïs dépassent aujourd'hui les dix milliards de dollars	*93* on dit ici que les avoirs du raïs dépasse aujourd'hui les dix milliards de dollars
94 et qu'à cela s'ajoutent encore des tonnes d'or entreposées dans les coffres helvétiques	*94* et qu' à cela s' ajoutent encore des tonnes d' or entreposées dans les coffres et le huit
95 et pour protéger cet empire une multitude de sociétés écrans autant de fiduciaires et l'assurance du secret bancaire	*95* qui pour protéger cet empire une multitude de sociétés écrans autant de fiduciaires et l' assurance du secret bancaire

96 tout ceci et puis au-dessus un homme vers lequel toutes les pistes convergent	*96* tout ceci et puis au dessus d' un homme vers lesquels toutes les pistes convergent
97 barzan al-tikriti demi-frère de saddam hussein un temps en poste à genève	*97* barzane al tikriti demi-frère de saddam hussein un temps en poste à genève
98 et dont le fils reste inscrit à l'université même s'il n'y a jamais mis les pieds	*98* et dont le fils reste inscrit à l' université même s' il n' y a jamais eu mille
99 al-tikriti le cerveau de la holding hussein souvent de passage à genève	*99* al tikriti le cerveau de la holding hussein souvent passage à genève
100 y compris ces derniers temps c'est d'ici qu'il a organisé le montage financier de la fortune accumulée	*100* y compris ces derniers temps c' est d' ici qu' il a organisé le montage financier de la fortune accumulée
101 son but blanchir l'argent de tous les trafics armes cigarettes et bien sûr pétrole	*101* son but de blanchir l' argent de tous les trafics armes cigarettes et bien sûr pétrole
102 du pétrole de contrebande qui passait les mailles du programme pétrole contre nourriture on dit que cette affaire-là	*102* du pétrole de contrebande qui passer les mailles du programme pétrole contre nourriture ont dit que cette affaire là
103 ramenait plus de deux milliards de dollars par an al-tikriti donc mais aussi un administrateur de société	*103* ramener plus de deux milliards de dollars par an à tikriti donc mais aussi un administrateur de sociétés
104 c'est à l'une d'entre elles qu'appartenait la montana-management	*104* c' était l' une d' entre elles qui appartenait la montana management
105 société immatriculée certes au panama mais appartenant bien à saddam hussein	*105* société immatriculée certes au panama mais appartenant bien à saddam hussein
106 des sociétés écrans pour une fortune colossale et répartie dans de multiples banques suisses	*106* des sociétés écrans pour une fortune colossale et répartis dans de multiples banques suisses
107 celles observées justement par washington genève jacques prévost france inter	*107* celle observée justement par washington de genève j' ai prévu en france
108 et outre les hommes de pouvoir les américains recherchent toujours des armes de destruction massive washington selon cnn	*108* mais outre les hommes du pouvoir les américains recherchent toujours des armes de destruction massive washington selon cnn
109 va envoyer dans ce but une nouvelle équipe d'un millier d'hommes des spécialistes militaires mais aussi des civils	*109* renvoyés dans ce but une nouvelle équipe d' un millier d' hommes des spécialistes militaires mais aussi des civils
110 kofi annan rappelle que seuls les inspecteurs des nations unies avaient l'autorité légale pour superviser le désarmement	*110* kofi annan rappelle que seuls les inspecteurs des nations unies avait l' autorité légale pour superviser le désarmement
111 hans blix évoquera une éventuelle reprise des inspections la semaine prochaine devant le conseil de sécurité	*111* ans blix évoquera une éventuelle reprise des inspections la semaine prochaine devant le conseil de sécurité
112 mais il souligne que les inspecteurs voudraient travailler en toute indépendance	*112* mais il souligne que les inspecteurs voudraient travailler en toute indépendance
113 nous ne serons pas des chiens au bout de la laisse américaine a laissé entendre le chef des inspecteurs en désarmement hier	*113* nous ne serons pas des chiens au bout de la laisse américaine a laissé entendre le chef des inspecteurs en désarmement hier
114 sur le terrain les américains se rendent compte au quotidien qu'il est décidément bien difficile de mettre de l'ordre	*114* sur le terrain les américains se rendre compte au quotidien qu' il est décidément bien difficile de mettre de l' ordre
115 illustration des déchirements de la population irakienne	*115* illustration des déchirements de la population irakienne
116 avec euh ce reportage de benoît collombat autour de l'hôtel palestine le lieu de résidence des journalistes	*116* avec euh ce reportage de benoît collombat autour de l' hôtel palestine le lieu de résidence des journalistes

117 de la presse internationale qui est aussi devenu le lieu des manifestations	*117* de la presse internationale qu' il cite aussi devenu le lieu des manifestations
118 les marines assistent à la scène médusés une manifestation de chiites avec slogans et banderoles	*118* les mêmes assisté à la scène médusé une manifestation le chili avec slogans et banderoles
119 nous voulons une véritable liberté le peuple irakien doit choisir ses gouvernants nous ne voulons pas les policiers de saddam	*119* nous voulons une véritable liberté le peuple irakien doit choisir ses gouvernants nous voulons pas les policiers de saddam
120 depuis quelques jours ils sont des centaines d'irakiens à se regrouper autour de l'hôtel palestine	*120* depuis quelques jours le sens des centaines d' irakiens se regrouper autour de l' hôtel palestine
121 transformé en bunker par l'armée américaine pour discuter protester revendiquer des débats souvent très vifs	*121* transformé en bunker par l' armée américaine pour discuter de protester revendiquer des débats souvent très vifs
122 moi je suis là parce que les américains sont en train d'occuper l'irak dit cet homme	*122* moi je suis là parce que les américains sont en train d' occuper l' irak la salle
123 mais dis-moi qu'est-ce que tu veux exactement l'interrompt un bagdadi dans la foule tu veux que saddam hussein revienne	*123* les dix mois qu' est ce que tu exactement la terre ont appris à faire du saddam hussein reviennent
124 qu'est-ce que tu proposes à la place des américains si tu as une meilleure idée je t'écoute ça m'intéresse	*124* qu' est ce que tu propose à l' appel des américains sont la meilleure et des écoles maternelles
125 donne-moi un nom comme président de l'irak eh bien ce sera le peuple répond le premier	*125* moins dont un président en irak mais ce sera le cas de le mettre en bière
126 le peuple oui c'est intéressant et concrètement tu fais comment	*126* le peuple russe est intéressant et concrètement physiquement
127 les éclats de voix se perdent dans la foule un homme est là avec la photo de ses frères	*127* les éclats de voix se perd dans la foule a nommé là avec la photo de ses frères
128 prisonniers politiques depuis onze ans la voix cassée d'avoir trop crié	*128* prisonniers politiques depuis onze ans la voix cassée d' avoir refusé
129 je suis allé voir les américains pour qu'ils m'aident à retrouver mes frères emprisonnés dans des cachots souterrains	*129* sur les américains qui n' a pas retrouvé les frères emprisonnés dans des cachots souterrains
130 les américains n'ont rien fait jusque-là	*130* les américains n' ont rien fait jusque là
131 un peu plus loin un ancien cadre du baas s'est reconverti en leader d'un nouveau parti l'union des irakiens	*131* un peu plus loin anciens cadres du baas et reconverti en leader d' un nouveau parti l' union des irakiens
132 ravi d'avoir retrouvé une tribune nous voulons que les forces américaines démilitarisent la ville	*132* a dit avoir retrouvé une tribu de trois nous voulons que les forces américaines démilitariser la ville
133 rétablissent l'eau et l'électricité ils ne font pas le nécessaire	*133* établi sur l' électricité ils ne sont pas nécessaires
134 et puis comme par défi un homme saute de trottoir en trottoir le drapeau irakien à la main	*134* et puis comme par des siens un homme sous de trottoir en trottoir le drapeau irakien à la main
135 ce drapeau je le porte très haut parce qu'il le mérite il ne faut pas le baisser tant qu'on est vivant	*135* drapeaux le porte très haut parce qu' ils le méritent qui ne faut pas le précédent en france
136 saddam hussein est le fils des américains les irakiens sont orgueilleux on n'est pas des lâches	*136* adam hussein le fils des américains les irakiens sont dans des lieux où des années là je
137 un habitant de saddam-city dit lui qu'il vient demander de l'aide aux américains	*137* un habitant de saddam city dix huit il vient demander de l' aide aux américains

138 il les avait applaudis quand ils sont entrés avec leurs chars mais aujourd'hui ils me déçoivent beaucoup	*138* les avaient applaudi quand ils sont entrés avec leurs chars mais aujourd'hui il ne déçoit beaucoup
139 bagdad éric damaggio benoît collombat france inter	*139* bagdad éric damaggio benoît collombat france-inter
140 enfin à propos de la reconstruction de l'irak washington a annoncé hier soir avoir euh attribué	*140* enfin à propos de la reconstruction de l' irak washington a annoncé hier soir avoir attribué
141 un gros contrat au groupe américain de bâtiment et de travaux publics begtel	*141* un gros contrat au groupe américain de bâtiment et de travaux publics becquerels
142 un contrat qui porte sur six cent quatre vingt millions de dollars	*142* un contrat qui porte sur six cent quatre-vingt millions de dollars
143 l'épidémie de pneumopathie atypique c'est une véritable course contre la montre	*143* l' épidémie de pneumopathie atypique c' est une véritable course contre la montre
144 qui est engagée pour le test de dépistage du virus	*144* s' est engagé pour le test de dépistage du virus
145 l'oms espère qu'il sera disponible avant trois semaines le secrétaire américain à la santé pense qu'il sera sur le marché	*145* l' oms espère qu' il sera disponible avant trois semaines le secrétaire américain à la santé pense qu' il sera sur le marché
146 d'ici dix jours	*146* d' ici dix jours
147 maigre butin hier lors de la fouille de la maison d'arrêt de luynes organisée après la triple évasion de lundi	*147* maigre butin hier lors de la fouille de la maison d' arrêt de luynes organisée après la triple évasion de lundi
148 deux portables quelques barrettes de shit et des armes artisanales à partir de fourchettes et de couteaux	*148* de portables quelques barrettes de shit que des armes artisanales à partir de fourchettes et couteaux
149 joey starr lui pourrait bien retourner en cellule le rappeur a écopé de quatre mois de prison dont un ferme	*149* des stars lui pourrait bien tourner en cellule rappeur écopé de quatre mois de prison dont un ferme
150 pour s'être euh rebellé et avoir craché sur des gendarmes mobiles	*150* pour s' être rebellés et avoir craché sur des gendarmes mobiles
151 joey starr qui en tant que vedette doit être un modèle social pour la jeunesse a rappelé le procureur	*151* les stars qui ont tant que vedette doit être un modèle social pour la jeunesse a rappelé le procureur
152 qui n'apprécie guère l'attitude provocante de l'artiste	*152* qui n' apprécient guère l' attitude provocante de l' artiste
153 de la provocation il y en avait mais c'était d'un tout autre artiste la collection surréaliste d'andré breton	*153* la provocation il en avait mais c' était un tout autre artiste la collection surréaliste d' andré breton
154 a donc été vendue aux enchères malgré la protestation d'intellectuels qui estiment qu'on a brisé l'oeuvre en la dispersant	*154* a donc été vendu aux enchères malgré la protestation d' intellectuels qui estime que l' on a brisé l' oeuvre en dispersant
155 cette vente s'est achevée hier soir les recettes ont largement dépassé l'estimation de départ	*155* cette vente s' est achevé hier soir les recettes ont largement dépassé les estimations de départ
156 une vente à l'image d'andré breton corinne audouin hors norme	*156* une vente à l' image d' ombre breton pour une fois hors normes
157 sur la tombe d'andré breton il est écrit je cherche l'or du temps	*157* sur la tombe d' andré breton il est écrit je cherche l' or du temps
158 l'or du temps c'est ce que l'écrivain avait amassé au fil des années dans son atelier de la rue fontaine	*158* lors du temps c' est ce que l' écrivain avait amassé au fil des années dans son atelier de la rue fontaine

159 si tout n'avait pas valeur de chef-d'oeuvre dans l'assemblage hétéroclite offert à la vente	*159* si tout n' avait pas valeur de chefs-d'oeuvre dans l' assemblage hétéroclite offerts à la vente
160 tout ou presque a suscité la passion des collectionneurs privés et de l'état	*160* tout ou presque a suscité la passion des collectionneurs privés et de l' état
161 qui se sont âprement disputés les lots les plus insignes	*161* qui se sont âprement disputés les eaux les plus assidus
162 parmi les trois cent trente cinq lots préemptés par l'état sur un total de quatre mille	*162* parmi les trois cent trente-cinq lots préempté par l' état sur le total de quatre mille
163 le manuscrit d'arcane dix sept de breton un relief en bois peint de jean arp	*163* le manuscrit de arcanes dix-sept de breton un relief en bois peint de jours par
164 plusieurs photos de man ray ont rejoint les collections publiques	*164* plusieurs photos de malheur est ont rejoint les collections publiques
165 deux tableaux phares de miro et magritte ainsi que le manuscrit de qu'est-ce que le surréalisme	*165* deux tableaux phare de minéraux et magritte ainsi que le manuscrit de casques le surréalisme
166 échoient en revanche à des collectionneurs privés	*166* échoit en revanche à des collectionneurs privés
167 tandis que des amateurs se sont arrachés les photomatons de breton et de ses amis surréalistes	*167* tandis que des amateurs se sont arrachés les photos matons de breton et de ses amis surréalistes
168 la cote de plusieurs peintres et de photographes s'est envolée du seul fait d'avoir appartenu à andré breton	*168* la cote de plusieurs peintres et photographes s' est envolée du seul fait d' avoir appartenu à andré breton
169 aux derniers jours de cette vente épique on peut se réjouir des passions que continue de déchaîner le surréalisme	*169* au dernier jour de cette vente épique on peut se réjouir des passions que continue de déchaîner le surréalisme
170 et aussi regretter que cette extraordinaire collection soit aujourd'hui éparpillée	*170* et aussi regretter que cette extraordinaire collection sont aujourd'hui éparpillés
171 chaque objet ayant perdu le caractère unique qu'il avait aux yeux du poète	*171* chaque objet ayant perdu le caractère unique qu' il avait su du poète
172 pour n'être plus réduit qu'à sa seule valeur marchande	*172* pour n' être plus réduits qu' à sa seule valeur marchande
173 souriez si vous prenez la route c'est la journée de la courtoisie au volant à l'occasion du week-end pascal	*173* sauriez si vous prenez la route c' est la journée de la courtoisie au volant l' occasion du week-end pascal
174 qui s'annonce très chargé rouge aujourd'hui et demain et lundi dans le sens des retours	*174* qui s' annoncent recharger rouge aujourd'hui et demain et lundi dans le sens des retours
175 souriez mais surtout soyez vigilants d'autant que les policiers eux sont de plus en plus vigilants	*175* selon les mais surtout soyez vigilants d' autant que les policiers ne sont de plus en plus vigilants
176 grâce à de nouveaux radars lancés dans le cadre de la lutte contre l'insécurité routière romain auzui	*176* grâce à de nouveaux radars lancer dans le cadre de la lutte contre l' insécurité routière en main aux oui
177 fini le temps des barbecues ces gros radars à lunettes qu'on peut voir sur les autoroutes bientôt tout sera automatisé	*177* fini le temps des barbecue ses gros radar à lunettes qu' on peut voir sur les autoroutes et bientôt tout sera automatisé
178 à la place du système magnétique qui vous flashe quand vous êtes en excès de vitesse	*178* à la place du système magnétique flash quand vous êtes en excès de vitesse
179 on aura droit à un radar numérique le but gagner en main-d'oeuvre plus besoin d'un gendarme qui prend la photo	*179* on aura droit un radar numérique le but de gagner en main-d'oeuvre plus besoin d' un gendarme qui prend la photo
180 mais surtout économiser du temps puisque le numérique qui transmet l'image quasiment en direct	*180* mais surtout économiser du temps puisque le numérique qui transmet les images quasiment en direct

181 remplacera le développement assez long des photos par conséquent les contrôles seront plus rapides et plus nombreux	*181* remplacera le développement assez long des photos par conséquent les contrôles seront plus rapide et plus nombreux
182 le fonctionnement est simple le radar sera installé sur une caméra numérique	*182* le fonctionnement est simple le radar sera installé sur une caméra numérique
183 donc quand vous serez en excès de vitesse celle-ci déclenchera automatiquement une impulsion	*183* donc en mousse en excès de vitesse celle ci déclenchera automatiquement une impulsion
184 et la photo sera transmise vers un ordinateur connecté à une sorte de grand fichier central des cartes grises	*184* et la photo soit transmis vers un ordinateur se connecter à une sorte de grand fichier central des cartes grises
185 à ce moment-là on pourra connaître le numéro de votre plaque d'immatriculation	*185* a souvent là on pourra connaître le numéro de votre pragmatique nation
186 et votre amende sera immédiatement imprimée puis envoyée	*186* et votre monde sera immédiatement imprimer puis envoyées
187 ces nouveaux radars commenceront à être mis en place dès la fin de l'année avec l'installation d'une centaine d'entre eux	*187* ces nouveaux radars commenceront à être mis en place dès la fin de l' année avec l' installation d' une centaine d' entre eux
188 l'objectif c'est qu'il y en ait mille en deux mille quatre essentiellement sur les autoroutes et les routes nationales	*188* l' objectif est étudiant les mines en deux mille quatre essentiellement sur les autoroutes et routes nationales
189 oui et levez le pied pour arriver euh tranquillement sur le sable par exemple	*189* oui et lever le pied pour aider euh tranquillement sur le sable par exemple
190 depuis quelques jours il fait aussi chaud sur la côte d'opale que sur la côte d'azur	*190* depuis quelques jours il fait aussi chaud sur la côte d' opale que sur la côte d' azur
191 mais l'eau est un petit peu plus fraîche comme l'a constaté laetitia charbonnel de france bleu nord	*191* mais l' eau étant de plus fraîche comme l' a constaté que laetitia charbonnel de france le nord
192 ah c'est gelé c'est très gelé l'eau	*192* à s' est jamais fait après les autres
193 en short et en tee-shirt stéphanie les pieds dans l'eau accompagne julie sa fille	*193* en short et un tee-shirt stéphanie les pieds dans l' eau accompagné jennifer tilly
194 qui n'hésite pas à se lancer tête la première dans les vagues	*194* qui n' hésite pas à se lancer tête la première dans les vagues
195 puis j'aime un petit peu bien l'eau froide alors euh me baigner tout le temps	*195* et italien lorsqu' elle est un peu limité état
196 alors euh en avril comme en juillet il faut prévoir la pelle le râteau et le seau	*196* à l' heure en avril comme en juillet il faut prévoir à l' appel de leurs actes et le saux
197 pour construire les incontournables châteaux	*197* pour construire les incontournables château
198 je vais prendre de l'eau dans les seaux et puis je vais jouer avec de l'eau euh dans le sable	*198* c' est vrai qu' un indice qui est un pays qui n' intéresse pas
199 c'est mieux l'été parce que l'été il fait très chaud	*199* c' est lui qui est pas un été sec et chaud
200 et après autant d'efforts le réconfort direction le marchand de glaces	*200* et après autant d' efforts le réconfort direction le marchand de glaces
201 paulette est derrière le comptoir	*201* paulette et derrière le comptoir
202 deux au chocolat et une à la	*202* nous sommes à la haine
203 bien une autre encore	*203* un autre plan
204 bon en général on vend un peu de tout hein chez nous on spécialisait surtout en croustillant hollandais	*204* en général on vint le tour chaîne officialiser surtout en caution demandée
205 donc c'est une affaire euh qui roule bien de ce côté-là	*205* non c' est une affaire euh qui au lieu de ce côté là

206 mais disons bon c'est vrai que les premières chaleurs comme ça c'est la glace hein	*206* mais disons bon c' est vrai que les premières chaleurs comme ça c' est la glace
207 à boulogne-sur-mer laetitia charbonnel france bleu nord pour france inter	*207* à bologne et sur mer avec ses charbonnel france bonheur pour france-inter
208 oui et sous le soleil de la côte d'azur il y a un joueur de tennis français qui est en train de se faire une place	*208* oui est sous le soleil de la côte d' azur veille un fort de tennis français qui est en train de se faire une place
209 quasiment inconnu du grand public julien boutter va jouer aujourd'hui en quart de finale du tournoi de monte-carlo	*209* quasiment inconnu du grand public julien boutter a joué aujourd'hui en quart de finale du tournoi de monte-carlo
210 face à l'espagnol carlos moya	*210* face à l' espagnol carlos moya
211 enfin divorce de jean tigana il a été démis de ses fonctions d'entraîneur dans le club anglais de fulham	*211* enfin le divorce de gens tigana il a été démis de ses fonctions d' entraîneur dans le club anglais de finale
212 qui appartient au richissime al fayed	*212* j' appartiens au richissime afrique
213 france inter l'actualité à huit heures un quart	*213* france-inter l' actualité huit heures un quart

Annexe E

Exemples de mots-clés

Afin de présenter les difficultés liées à la sélection de mots-clés à partir de sections thématiques, nous donnons deux extraits d'émission, qui ont été transcrits par le système de RAP Irene et dont les frontières ont été déterminées manuellement. Les termes représentatifs ont été obtenus en supprimant les mots grammaticaux et en recourant au *stemming*. Le critère employé ici pour choisir ces mots-clés est un critère *tf-idf* classique, dans le but de montrer l'intérêt des modifications que nous avons dues faire pour pénaliser les noms propres et les mots associés à un faible indice de confiance.

E.1 Premier extrait

Le premier segment sélectionné traite de la mise en place du « plan cancer ». Il est constitué d'une présentation du sujet par un journaliste, suivi d'une interview d'une personne chargée de la psycho-oncologie à la Ligue nationale contre le cancer (Fig. E.1). Si l'on restreint la liste des mots à sept, les mots-clés proposés sont « *cancer* », « *psychopathologie* », « *malades* », « *psychologique* », « *cancérologue* », « *cancéreuses* » et « *psychologue* ». Les termes proposés sont très satisfaisants puisqu'ils caractérisent très bien le sujet abordé tout en retournant un nombre significatif de réponses de la part du moteur de recherche. Ils illustrent en outre l'intérêt de sélectionner directement des informations à partir du texte plutôt que de recourir à un ensemble de thèmes fixé *a priori* puisqu'il est possible que des sujets liés au cancer et à la psychologie ne figurent pas dans cette liste.

E.2 Second extrait

Dans un second texte choisi comme contre-exemple du précédent, un journaliste consacre une page sportive au licenciement de l'entraîneur de football Jean Tigana, à l'actualité du championnat français de football, au joueur de football Marcel Desailly, aux prochains championnats du monde d'athlétisme et au tournoi de tennis de Monte-Carlo (Fig. E.2). Cette section montre une des principales difficultés pour établir une segmentation thématique d'une émission d'information puisque suivant la granularité

il y a presque un an lorsque le président de la république // a annoncé // c' est moi // presque un génie un an // cinq mois // lorsque le président de la république annonçait ce qu' on appelle maintenant le plan cancer // euh il a évoqué pour tous les malades une prise en charge globale c'est-à-dire médicale sociale // et psychologique // la psychopathologie et spécialité récente puisque le manuel qui devait signer son acte de naissance est paru en mille neuf cent quatre-vingt-neuf // la psychopathologie // c' est donc là de quelques six que le web s' y tient // peut apporter à malades cancéreux // qui doit faire face à sa maladie et à une médecine très technique // qui ne répond // ni à ses angoisses ni à sa peur de mourir // rendons à césar ce qui est assez un // avant mille neuf cent soixante // le professeur jean-bernard qui soigner les enfants le scénique // avait déjà demandé à une psychologue // nicolas libye // de s' occuper de ses enfants // aujourd'hui nicolas libye est chargé de la psychologie de la ligue contre le cancer // et dix ans après par le docteur daniel sera cancérologue de son état le président de la société française de psychopathologie // les farouches défenseurs de cette spécialité // à qui s' adresse la psychopathologie // aux malades // à leurs familles // et aux soignants // pas à tous les malades certains n' ont pas besoin // mais les autres // au moment clé de la maladie // à l' annonce du diagnostic de cancer // retentissement psychologique inévitable // lors des traitements lourds et pénibles // fatigue perte de cheveux // et à l' arrêt du traitement // quand il faut renouer avec la vie professionnelle et sociale // en outre de graves problèmes // le cancer // c' est la pathologie de la certitude // on n' est jamais sûr d' être guéris // et cette maladie pour plus de la moitié des patients // à devenir chronique // et s' étendre sur des années // lorsque quatre-vingt-dix-huit la ligue contre le cancer avait réuni à paris // mille malades // première chose qu' ils avaient tous demandé // plus de psychologie s' il vous plaît // et une aide quand le besoin s' en fait sentir // les cancérologue en ont été saisis // il n' avait pas mesuré l' importance de l' écoute // de la prise en charge psychologique // trop occupé // par le combat à mener contre les cellules cancéreuses // les malades ont donc besoin de s' exprimer // le psychologue est alors le réceptacle des angoisses // et le psy aide à se reconstruire // aide à reconstituer l' équilibre constamment menacées // par les aléas de la maladie // donc là aussi qu' on oncologie pas pour tous // mais pour beaucoup // pour les jeunes gens atteints du cancer existe une association qui s' appelle ici est c' est c' est ce qui signifie jeune solidarité cancer // et un site internet dans la cgc forum tous attachés j' ai ces forums point net // là on peut échanger les expériences connaître les droits // et puis surtout fait sentir moins seul

FIG. E.1: *Transcription automatique de la partie d'une émission traitant du « plan cancer », où // marque une fin de groupe de souffle*

jan tigana remercie de flamme le club de foot euh anglais les résultats obtenus par l' entraîneur français n' étant pas à la hauteur des ambitions du club // dans le championnat de ligue un ça va être euh le sprint pour les meilleurs ce week-end germain treille // à quatre journées de la fin c' est le moment pour les équipes qui peuvent prétendre au titre de se démarquer // avant cette euh trente-quatrième journée lyon et monaco sont les deux formations les mieux placés pour le titre // marseille et bordeaux ne sont pas écartés de cette course mais sont distancés avec respectivement quatre et six points de retard sur l' om // d' autant que ce week-end lyonnais et monégasques en pleine confiance en ce moment l' occasion de creuser l' écart avec leurs rivaux // et ce avec l' avantage de recevoir lyon contre l' idée monaco face à strasbourg // car juste derrière marseille qui accumule les revers se déplacent à lens // et bordeaux va jouer à ajaccio qui lutte pour le maintien // hier marcel desailly le capitaine de l' équipe de france de football et de passage à paris pour faire la promotion d' un ouvrage consacré au sport spectacle // à cette occasion il s' est défendu d' être un homme d' affaires // et dix mois après l' échec du mondial deux mille deux de ses yeux reconnaît qu' il n' est plus sollicité de la même manière // peut-être n' est il pas payer les factures de six un téléphone portable // le pavillon belge de la cité universitaire de paris deux cents chambres individuelles et quelques studios // c' est là que sera logé fin août l' équipe de france de philippe de athlétisme pendant les championnats du monde // et un groupe de soulagement ils pourront manger ailleurs qu' aux restos // et julien boutter les pieds sur terre battue mais la tête dans les nuages heureux le jeunisme français encore en lice d' un tennis de monte-carlo // quoi qu' il arrive dans son match de quarts de finale face à carlos moya finaliste l' an dernier // ce sera pour lui des bonus

FIG. E.2: *Transcription automatique de la partie d'une émission correspondant à une chronique sportive, où // marque une fin de groupe de souffle*

choisie sur ce que représente un thème, cet extrait peut constituer un tout lié au sport, ou bien être divisé en cinq puisque cinq faits principaux sont décrits ici ou même en trois si on retient comme critère de division le sport traité. Les mots-clés retenus par le critère *tf-idf* usuel pour l'ensemble de cette section sont cette fois-ci « *écart* », « *bordeaux* », « *lyon* », « *équipe* », « *jeunisme* », « *club* » et « *démarquer* ». À la lecture de cette liste, il est difficile de se faire une idée du thème traité puisque seuls « *équipe* » et « *club* » sont susceptibles d'être liés au sport. Les villes de « *bordeaux* » et « *lyon* » sont par exemple anecdotiques par rapport aux sujets abordés, mais pourtant sélectionnés du fait de leur rareté relative. Le terme « *jeunisme* » a quant à lui été transcrit de manière incorrecte par le système de RAP, en remplacement de la séquence « *jeune et seul* » réellement prononcée. Les modifications que nous proposons du critère *tf-idf* visent à éviter de sélectionner de tels mots-clés, en favorisant d'autres termes comme « *championnat* ».

Résumé

Une des voies pour améliorer les performances des systèmes de reconnaissance automatique de la parole (RAP) consiste à employer davantage de connaissances linguistiques. Nous nous plaçons dans ce cadre en axant nos travaux sur deux aspects : les informations morpho-syntaxiques et l'adaptation thématique.

Dans une première partie, nous proposons une nouvelle prise en compte des parties du discours en post-traitement du décodage de la parole. Nous étiquetons pour ce faire des listes d'hypothèses à l'aide d'un analyseur morpho-syntaxique développé pour tenir compte des spécificités des transcriptions. Nous réordonnons ces listes en modifiant le score global employé par un système de RAP pour reconnaître un groupe de souffle, de manière à inclure les informations morpho-syntaxiques. Une diminution significative du taux d'erreur de mots est obtenue lors d'expériences menées sur des émissions d'information francophones (corpus ESTER). Nous établissons en outre l'intérêt des informations morpho-syntaxiques pour améliorer des mesures de confiance.

Dans une seconde partie plus exploratoire, nous nous intéressons à l'adaptation thématique d'un modèle de langage (ML) d'un système de RAP. Nous proposons à cette fin une chaîne de traitements originale qui nous permet de spécialiser de manière fine et non supervisée le ML. Nous découpons tout d'abord une première transcription du document étudié en sections thématiquement homogènes. Pour cela, nous développons un nouveau cadre probabiliste pour prendre en compte différentes modalités (cohérence lexicale, indices acoustiques et marqueurs linguistiques) et montrons son intérêt pour améliorer la qualité de la segmentation. Nous construisons ensuite des corpus d'adaptation à partir du Web en utilisant une procédure novatrice. Nous modifions enfin le ML avec ces corpus spécifiques et montrons que, sur des sections thématiques sélectionnées manuellement, cette méthode permet d'améliorer significativement la modélisation du ML, même si le gain constaté au niveau de la qualité de la transcription est léger.

Abstract

A way to improve outputs produced by automatic speech recognition (ASR) systems is to integrate additional linguistic knowledge. Our research in this field focuses on two aspects: morpho-syntactic information and thematic adaptation.

In the first part, we propose a new mode of integration of parts of speech in a post-processing stage of speech decoding. To do this, we tag N-best sentence hypothesis lists with a morpho-syntactic tagger built to take into account the specificities of transcriptions. We reorder these lists by modifying the score computed by an ASR system at the sentence level to include morpho-syntactic information. Experiments done on French-speaking broadcast news (ESTER corpus) exhibit a significant improvement of the word error rate. Besides, we establish the contribution of morpho-syntactic information to improve posterior based confidence measures.

In the second more exploratory part, we are interested in thematically adapting the language model (LM) of an ASR system. We propose a scheme that enables us to specialize speech decoding in an unsupervised way. We first segment the studied document into thematically homogeneous sections. To this end, we develop a new probabilistic framework to integrate different modalities (lexical cohesion, acoustic clues, and linguistic markers) and show its relevance to improve segmentation. We then build adaptation corpora retrieved from the Web by using an innovative procedure. We finally modify the LM with these specific corpora and show that, on thematic sections that are manually selected, this method significantly improves the LM, even if the increase of the word error rate is slight.