Modélisation linguistique pour l'indexation automatique de documents audiovisuels

Alexandre Allauzen
Thèse effectuée au LIMSI-CNRS sous la direction de Jean-Luc Gauvain,
en collaboration avec l'Institut National de l'Audiovisuel (Laurent Vinet)

Le jury

Thèse soutenue le 23 décembre 2003 devant la commision d'examen :

Régine André-Obrecht	IRIT, Toulouse	Rapporteur
Marc El-Bèze	LIA, Avignon	Rapporteur
Jean-Luc Gauvain	LIMSI-CNRS, Orsay	Directeur
Edouard Geoffrois	DGA-CTA, Arcueil	Examinateur
Laurent Vinet	INA, Bry-sur-Marne	Examinateur

Remerciements

Cette thèse s'est déroulée au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), au sein du groupe Traitement du Langage Parlé (TLP) qui s'est illustré par sa cohésion et sa solidarité. Je remercie mon directeur de thèse. Il définit bien ce que, je pense, doit être un directeur de thèse, tant par son enseignement, que par sa motivation à faire de la recherche. Je tiens à remercier également Gilles Adda pour sa patience puis pour ses relectures et ses conseils si nombreux, Patrick Paroubek pour avoir été bien plus qu'un collègue de bureau, Lori Lamel et Claude Barras qui m'ont initié aux joies de l'écriture d'articles, Eric Bilinski à qui je dois beaucoup techniquement, Sophie Rosset pour les ressources et les conseils si précieux, Kevin Mc Tait pour les café du matin, et tout le groupe TLP pour m'avoir aidé jusqu'à la soutenance. Le LIMSI est un lieu d'accueil privilégié, et j'en remercie ses directeurs Joseph Mariani puis Patrick Le Quéré. Jamal Atif et Xavier Ripoche ont participé de manière décisive à ma thèse et je les en remercie chaleureusement. Je n'oublie pas Romain Prudon et Frederic Joly pour leur soutient si précieux.

Je voudrais ensuite remercier mon jury : mes deux rapporteurs, Régine André-Obrecht et Marc El-Bèze, bien sûr Jean-Luc Gauvain, Edouard Geoffrois et Laurent Vinet, qui ont accompli un travail de relecture minutieux et constructif. Accepter de lire et de critiquer les travaux d'un doctorant, c'est d'abord accepter une charge de travail conséquente. Les membres de mon jury ont fait bien plus, puisqu'ils ont amélioré et enrichit mon manuscrit.

Au départ de cette thèse, il y a eu un stage de DEA à la Direction de la Recherche et de l'Expérimentation de l'Institut National de l'Audiovisuel qui s'est poursuivit par une collaboration enrichissante lors de ma thèse. Je remercie ses directeurs Bruno Bachimont et Daniel Terrugi pour leur accueil et soutient. Laurent Vinet, qui m'a encadré, et Emmanuel Veneau m'ont tout les deux beaucoup apporté, et c'est un doux euphémisme. Je les remercie chaleureusement ainsi que Rémi Landais et sa jeunesse impétueuse. C'est grâce à l'INA et plus particulièrement à Michèle Wautelet, que j'ai découvert la pratique documentaire.

Enfin, je tiens à remercier Maxence Klein, il sait pourquoi. Et Patricia qui m'a supporté et sans qui, il n'y aurait rien.

Sommaire

Le jury iii

Remerciements v

Introduction xv

1 Pratique documentaire audiovisuelle **1**

1.1 Définition de l'indexation du point de vue de la documentation . . . 1

1.2 Spécificité de l'indexation des documents audiovisuels . . . 2

1.2.1 Principales spécificités du document audiovisuel . . . 2

1.2.2 Une indexation nécessairement subjective . . . 3

1.2.3 Une indexation guidée par l'usage . . . 3

1.3 Pratique documentaire audiovisuelle : l'INA . . . 4

1.3.1 Missions et quantité . . . 4

1.3.2 Documentation d'une émission . . . 4

1.3.3 De l'analogique vers le numérique . . . 6

1.3.4 Usages de l'indexation . . . 6

2 Indexation automatique de documents audiovisuels **9**

2.1 Indexation manuelle *vs* automatique . . . 9

2.1.1 Automatiser l'indexation manuelle ou indexation automatique . . . 9

2.1.2 Le passage au sémantique . . . 10

2.2 Indexation automatique de document audio . . . 11

2.2.1 Motivation . . . 11

2.2.2 Système automatique d'indexation de documents audio . . . 12

2.2.3 Spécificité de la parole transcrite 13

2.2.4 Évaluation d'un système *SDR* 13

2.2.5 Effet des erreurs de reconnaissance 15

2.3 Transcription enrichie et indexation vidéo 16

2.3.1 Transcription Enrichie 16

2.3.2 Indexation vidéo 17

2.4 Conclusion et position de la thèse 18

3 Reconnaissance automatique de la parole **21**

3.1 Reconnaissance statistique de la parole 21

3.1.1 Codage du signal 22

3.1.2 Modélisation acoustique 24

3.1.3 Lexique de reconnaissance 24

3.1.4 Modélisation linguistique 25

3.1.5 Décodeur 26

3.2 Système de transcription automatique 27

3.2.1 Le partitionneur 27

3.2.2 Taux d'erreur sur les mots 27

4 Modélisation statistique du langage **31**

4.1 Modélisation statistique du langage 32

4.1.1 Le modèle *n-gram* de mots 33

4.1.2 Le modèle *n-gram* de classes 35

4.1.3 Modèle connexioniste 36

4.1.4 Autres modèles 37

4.1.5 Faiblesses de la modélisation statistique du langage 37

4.2 Combinaison de modèles et introduction de contraintes 40

4.2.1 Interpolation linéaire 40

4.2.2 Maximum d'entropie 41

4.3 Évaluation des modèles de langages 41

4.3.1 Perplexité 42

4.3.2 Perplexité *vs* WER 43

4.4 Construction du modèle *n-gram* de référence 43
4.4.1 Description des corpus 44
4.4.2 Définition d'un mot et normalisation des textes 45
4.4.3 Estimation des probabilités 47
4.4.4 Lissage ou *smoothing* 48
4.4.5 Utilisation des corpus 50
4.4.6 Description du modèle de langage de référence 50
4.5 Adaptation des modèles de langage 50
4.5.1 Les Modèles adaptatifs 51
4.5.2 Adaptation non supervisée des modèles de langage 53
4.5.3 Adaptation utilisant des connaissances *a priori* 53

5 Construction automatique du vocabulaire 59
5.1 Cadre d'évaluation 60
5.1.1 Corpus d'entraînement 60
5.1.2 Critères d'évaluation 60
5.1.3 Le vocabulaire idéal 61
5.2 Le vocabulaire de référence 62
5.2.1 Cadre théorique 63
5.2.2 Construction du vocabulaire de référence 63
5.2.3 Typologie des mots hors vocabulaire 64
5.3 Construction automatique du vocabulaire 65
5.3.1 Segmentation des corpus d'entraînement 65
5.3.2 Choix des vocabulaires pour l'estimation des distributions 66
5.3.3 Opérateur et fonction coût 67
5.3.4 Algorithme de construction automatique du vocabulaire 68
5.4 Expérimentations 68
5.4.1 Validation 69
5.4.2 Segmentation par années et par sources 69
5.4.3 Segmentation hybride 70
5.4.4 Équivalence entre perplexité et couverture lexicale 71

6 Adaptation du vocabulaire 75

6.1 Corpus extrait du Web . . . 76

6.1.1 Normalisation des textes . . . 77

6.1.2 Constitution des corpus . . . 77

6.2 Vocabulaire de référence . . . 78

6.3 Adaptation *ad hoc* . . . 79

6.3.1 Heuristique pour la mise à jour du vocabulaire . . . 79

6.3.2 Expérimentation . . . 80

6.3.3 Analyse quantitative des mots ajoutés . . . 81

6.3.4 Pérennité de l'adaptation *ad hoc* . . . 81

6.3.5 Faiblesses de l'algorithme . . . 83

6.4 Adaptation par interpolation linéaire de modèle unigramme . . . 83

6.4.1 Algorithme d'adaptation du vocabulaire . . . 83

6.4.2 Expérimentation . . . 84

6.5 Adaptation vectorielle du vocabulaire . . . 85

6.5.1 Description de l'algorithme vectoriel . . . 85

6.5.2 Expérimentation . . . 86

6.5.3 Analyse quantitative des mots ajoutés . . . 86

6.5.4 Pérennité de l'adaptation vectorielle . . . 88

6.6 Comparaison des algorithmes *ad hoc* et vectoriel . . . 88

6.7 Dimension du vocabulaire . . . 89

7 Construction d'un modèle de langage à vocabulaire ouvert 93

7.1 Repli lexical . . . 94

7.1.1 Choix des formes de repli . . . 94

7.1.2 Construction du modèle de langage . . . 95

7.1.3 Introduction de mots nouveaux et phonétisation . . . 96

7.2 Expérimentation . . . 96

7.2.1 Corpus de test . . . 97

7.2.2 Expérimentation interactive . . . 97

7.2.3 Expérimentation automatique . . . 99

8 Adaptation automatique du modèle de langage 103

8.1 Transcription automatique d'archives audiovisuelles 104

8.1.1 Description du corpus 104

8.1.2 Analyse des erreurs du système standard 105

8.1.3 Adaptation des modèles de langage 106

8.2 Transcription automatique d'émission d'actualité 111

8.2.1 Corpus de test et d'adaptation 111

8.2.2 Interpolation linéaire 112

8.2.3 Interpolation sélective 113

8.2.4 Expérimentation sur les émissions radiodiffusées 113

8.2.5 Expérimentation sur les émissions télédiffusées 116

8.2.6 Expérimentation EuroNews 118

Résumé

La plupart des méthodes d'indexation de documents audiovisuels sont en pratique manuelles. En France, l'Institut national de l'audiovisuel (INA) archive et indexe tous les jours plus de 20 chaînes de radio ou de télévision, ce qui représente plus de 100 000 heures par an auxquelles s'ajoutent plus d'un million d'heures déjà archivées. L'inclusion de traitements automatiques dans le processus d'indexation doit donc être pensé afin de répondre aux usages réels d'un institut comme l'INA, et aux spécificités de l'objet audiovisuel. L'état de l'art en matière d'indexation automatique de documents audiovisuels est constitué d'un système de reconnaissance automatique de la parole allié à des techniques de recherches d'information. La transcription automatique de la bande sonore est donc l'accès premier au contenu audiovisuel et les erreurs qu'elle peut contenir en détermine la pertinence. Les erreurs de transcription sont principalement dues à l'écart entre ces modèles et la réalité rencontrée dans les documents audiovisuels. En particulier, le contenu lexical et linguistique des transcriptions automatiques est conditionné par le vocabulaire et le modèle de langage du système.

Cette thèse porte sur l'adaptation du vocabulaire et du modèle de langage d'un système de transcription automatique en vue de l'indexation de documents audiovisuels qu'ils soient issus d'archives, ou collectés quotidiennement. Dans les deux cas, une des difficultés est de disposer de données au format électronique, qui soient appropriées et contemporaines des documents à transcrire. La première solution explorée dans cette thèse est de construire un modèle de langage dont le vocabulaire est ouvert, et qui peut être modifié de manière interactive. La seconde consiste à utiliser des données provenant de sites internet pour constituer les ressources textuelles pertinentes et contemporaines de l'émission. La solution proposée pour construire un modèle de langage à vocabulaire ouvert s'appuie sur la technique du repli lexical. Ce modèle est expérimenté sur un corpus de journaux télévisés. Lorsque les mots hors vocabulaire nouveaux sont introduits manuellement, près de 80% de leurs occurrences sont reconnus. Cela permet une réduction du taux d'erreur sur le mots de 2.6% absolu.

Les expériences sur les documents d'archive utilisent le corpus ECHO qui couvre une période allant des années 40 à 90. Les données d'apprentissage du modèle de langage datent principalement des années 90. Cette différence d'époque se traduit par une augmentation de la perplexité pouvant aller jusqu'à 40%. Un modèle de langage d'époque est estimé à partir de textes anciens collectés sur des sites internet. Le modèle de référence est alors adapté en l'interpolant avec ce modèle. Grâce à cette méthode, la réduction de la perplexité observée sur les documents des années 40 est de 9% et décroît jusqu'à s'annuler pour les années 90. Ces gains en perplexité n'impliquent pas de réduction significative du taux d'erreur sur les mots. Pour la transcription automatique des journaux télévisés diffusés tous les jours, une expérimentation quotidienne est effectuée sur le mois de janvier 2002. L'objectif est de développer des méthodes originales permettant d'adapter le vocabulaire et les modèles de langage à partir de données d'adaptation collectées le jour même sur des sites internet. Afin d'étudier le décalage entre le corpus d'adaptation et le corpus de test, l'expérience est menée sur trois corpus de test provenant de média distincts. Suivant les cas, les techniques développées permettent un gain relatif pouvant aller jusqu'à 50% sur la couverture lexicale, et jusqu'à 25% sur les performances de reconnaissance.

Introduction

« Wait a minute, wait a minute. You ain't heard nothin' yet ! » [1]. Le premier document audiovisuel parlant s'ouvrait sur une phrase prémonitoire, et la quantité de documents audiovisuels ne cesse effectivement de croître. Cette croissance est d'autant plus rapide que les documents sont faciles à produire et à diffuser. Parallèlement, les développement des dispositifs de numérisation et de stockage permettent de conserver toutes les données désirées pour constituer des « bibliothèques audiovisuelles ». Mais ces capacités d'archivage numérique ne doivent pas faire oublier l'avertissement affiché dans toutes les bibliothèques : un document mal rangé est un document perdu. Dans le cadre d'une bibliothèque audiovisuelle, cette règle d'or soulève les questions essentielles de comment classer un document audiovisuel, et avant tout, comment l'indexer. En France, l'Institut National de l'Audiovisuel (INA) a en charge la gestion du patrimoine audiovisuel français. Cela se traduit par des quantités édifiantes : un peu plus de 1,5 millions d'heures sont stockés dans plus d'une centaine de kilomètres de rayonnage, et chaque année s'ajoutent près de 100 000 heures. Ces ordres de grandeurs révèlent la masse des documents collectés, l'étendue du travail d'indexation qui est aujourd'hui manuel, et donc la nécessité d'une solution automatique à l'indexation des documents audiovisuels.

Indexation automatique de documents audiovisuels

L'indexation est historiquement une pratique documentaire dédiée aux textes, et dont l'objectif est de générer pour un document, une description sous forme d'index permettant de le retrouver dans un ensemble. Le choix des caractéristiques du document et du langage de description détermine la pertinence des index. Ces questions sont au centre des préoccupations de l'INA, mais également des recherches actuelles sur l'indexation automatique de documents audiovisuels qui sont en grande partie fédérées par les campagnes d'évaluation du National Institut of Standardization and Normalization (NIST). Mais si ces questions sont les mêmes, les approches adoptées et les pratiques qui y sont associées diffèrent. Un institut tel que l'INA a en charge l'indexation des archives audiovisuelles française et doit répondre aux requêtes de ces clients (professionnels de l'audiovisuel, chercheurs). Il possède donc la double expérience de l'indexation

[1] C'est avec 281 mots que le parlant fit son entrée dans l'histoire du cinéma le 6 octobre 1927 au Warner Theater de New York. Le film « The Jazz Singer » et son célèbre "Attendez un peu, vous n'avez encore rien entendu" entrait dans la légende dès sa première représentation publique grâce à une partie sonore enregistrée sur disque, le système Vitaphone.

et de l'usage de cette indexation. Ainsi l'INA a défini ses propres méthodes et son propre langage de description. L'élément central en est la notice documentaire. Elle est construite manuellement par une équipe de documentalistes et regroupe l'ensemble des index. Cette notice se substitue au document lors de la phase de recherche.

La dernière campagne d'évaluation TRECVID[2] donne un aperçu des recherches en indexation automatique de documents audiovisuels et des systèmes mis en œuvre. Ses conclusions mettent en exergue la difficulté à extraire de l'image des caractéristiques (comme la détection des visages ou des changements de plan et les textures) exploitables en vue de l'indexation d'un document audiovisuel [Quénot *et al.*, 2002]. Les systèmes présentés qui se fondent uniquement sur les techniques d'analyse automatique de l'image obtiennent des résultats peu convaincants. Si l'utilisation conjointe de ces caractéristiques avec celles extraites de l'audio (principalement la transcription automatique et la segmentation acoustique) aboutit à une indexation plus satisfaisante, les systèmes les plus performants n'utilisent que les résultats de la transcription automatique de la parole. Ainsi, pour comparer l'approche manuelle à une approche réaliste en indexation automatique, il faut se tourner vers une approche de type *SDR* (*Spoken Document Retrieval*), approche validée lors des précédentes évaluations [Garofolo *et al.*, 1999]. Ce type de système allie les caractéristiques qui sont extraites de la transcription automatique de ce qui est dit, à un moteur de recherche.

Afin de confronter les deux approches de l'indexation, le tableau 1 compare d'une part, un extrait du champ « Résumé »[3] d'une notice documentaire, et d'autre part, la transcription automatique du passage correspondant. La similarité de contenu entre les deux est flagrante. La transcription donne même un complément d'information par rapport à la notice sur la position du Ministre de l'époque. Des quelques erreurs de transcriptions, seules deux peuvent être pénalisantes dans une perspective d'indexation. La première erreur, sur le mot « totalité », modifie le sens de la phrase. La seconde porte sur le mot « nationalité », et bien que ce mot soit absent de la notice, il est susceptible de faire partie d'une requête.

Transcription automatique de documents audiovisuels

Les performances d'un système de type *SDR* sont dépendantes du contenu des transcriptions automatiques, qui n'est pas toujours aussi favorable à l'indexation que celui présenté dans le tableau 1. Comme le montrent les exemples de transcription du tableau 2, les erreurs présentes dans la transcription automatique n'ont pas toutes les mêmes incidences dans une perspective d'indexation. Par exemple, la suppression de la négation « n' » par le système de reconnaissance n'a aucune importance dans la cas qui nous concerne. Les erreurs sur les noms propres comme

[2]TRECVID est la contraction de « TREC Video ». Ces évaluations, et les conférences qui y sont associées, concernent spécifiquement l'indexation de documents audiovisuels et elles s'inscrivent dans le cadre plus large des évaluations TREC (Text REtrival Conferrence) qui portent sur l'indexation en général. Il est fait référence ici à la conférence TRECVID 2002, la conférence TRECVID 2003 n'ayant pas encore eu lieu au moment où ces lignes sont écrites.

[3]Dans le cadre d'une recherche par le contenu, le champ « Résumé » et le champ « Descripteurs » sont les deux permettant de répondre à une requête. Les autres champs correspondent à des données techniques de catalogage.

Notice documentaire	Transcription automatique
TITRE :Manifestation sans-papier	une manifestation des sans-papiers un an après Saint-Bernard (...)
EMIGRATION : Un an après l'évacuation de l'église Saint-Bernard, environ 2 000 personnes ont manifesté cet après midi-midi pour demander l'abrogation des lois PASQUA-DEBRE et pour dénoncer la décision du gouvernement de ne pas les supprimer en **totalité**.	un an après l'évacuation de l'église Saint-Bernard à Paris environ deux mille personnes ont manifesté cet après-midi un rassemblement pour demander l'abrogation des lois Pasqua Debré beaucoup ont dénoncé la décision du gouvernement de ne pas les supprimer en **Bretagne il y a à** cet après-midi Elisabeth Guigou ministre la justice a affirmé que le droit de décider s' il serait bien **réformes**

TAB. 1 – Pour le passage d'un journal télévisé, ce tableau compare l'extrait de la notice documentaire de l'INA et le passage de la transcription automatique correspondant. Le texte en commun est en noir. Dans la transcription, le mot en gras « Bretagne » correspond à une erreur du système de reconnaissance, il aurait dû correspondre dans la notice au mot « totalité », également entouré. Les autres erreurs sont en gras.

la substitution de « Paris » par « pari », ou l'ambiguïté entre « Gréco » (Juliette) et le peintre « le Greco », peuvent elles porter à conséquence suivant que les recherches ultérieures portent sur la peinture, Saint-Germain-des-Prés, ou la ville de Paris.

Le système de reconnaissance automatique de la parole utilise pour générer ces transcriptions des sources de connaissances qui tentent de capturer les caractéristiques acoustiques, lexicales et linguistiques du langage parlé au travers des modèles acoustiques, du vocabulaire, du dictionnaire de prononciation, et du modèle de langage [de Mori and Bigi, 2002; Gauvain and Lamel, 2002]. Les erreurs de transcription sont principalement dues à l'écart entre ces modèles et la réalité rencontrée dans les documents audiovisuels. En particulier, le contenu lexical et linguistique des transcriptions automatiques est conditionné par le vocabulaire[4] qui est de taille limitée (en général inférieur à 65 535), et le modèle de langage[5]. Les extraits de transcriptions du tableau 2 illustrent l'impact de ces deux composantes. Dans le premier extrait, que ce soit « Paris » et « pari », ou « Gréco » et « Greco », tous ces mots sont dans le vocabulaire et homophones et les erreurs sont donc dues au modèle de langage. Dans le second extrait, le mot « Alphaville »[6] est absent du vocabulaire, il ne peut donc être reconnu et cause des erreurs dans son voisinage.

La modélisation lexicale et linguistique utilisée par le système de reconnaissance se fonde sur les observations faites sur des corpus d'entraînement. Il est donc nécessaire que des connaissances sur l'émission à transcrire soient contenues dans ces corpus, pour que les modèles soient pertinents. Mais étant donné le coût important lié à leur préparation, les données d'apprentissage disponibles sont en général figées dans le temps. Pour un système de transcription automatique de

[4]Le vocabulaire est la liste close des mots pouvant être reconnus par le système de transcription.
[5]Le modèle de langage assigne une probabilité à chaque suite de mots extraite du vocabulaire
[6]Ville imaginaire où se déroule le film de J-L. Godard.

Transcription automatique	Référence
le **pari** de Saint-Germain-des-Prés et pas seulement celui **du Greco** du be-bop c' est aussi celui des affrontements entre les intellectuels que la guerre froide divise profondément enfin les acteurs de ce débat sont Jean Paul Sartre **Maurice Peyrot continue** Albert Camus et Raymond Aron	le **Paris** de Saint-Germain-des-Prés **n'** est pas seulement celui **de Gréco** du be-bop c' est aussi celui des affrontements entre les intellectuels que la guerre froide divise profondément enfin les acteurs de ce débat sont Jean Paul Sartre **Maurice Merleau Ponty** Albert Camus et Raymond Aron
quatre cent mille habitants en **dépit** peut être à mi chemin entre **la ville voisine** que sera **sa**	quatre cent mille habitants en **l' an deux mille** peut être à mi chemin entre **Alphaville ou Brasilla** que sera **Cergy**

TAB. 2 – Deux extraits de transcription d'émission d'archive. Le premier date de 1970 et le second de 1981. Le texte en gras souligne les erreurs du système de reconnaissance.

documents audiovisuels, on utilise des textes de journaux et de transcriptions manuelles datant principalement des années 90. Ce type de données ne contiennent pas l'information suffisente pour faire face à la variété rencontrée dans les émissions radio et télévisées. Cette variété se décline à la fois en terme d'époque et de type de documents.

Prenons comme exemple les émissions d'archive. Entre les actualités de l'après guerre et les premiers journaux télévisés des années 60-70, le contenu linguistique a évolué, et avec lui le style de parole et les thèmes abordés (du plan Marshall à l'Amoco Cadiz). De plus, les progrès techniques ont permis une prise de liberté de plus en plus grande dans la forme même des émissions avec l'apparition des interviews et des séquences extérieures. Il n'y a pas une réelle adéquation entre ce type de documents et les données textuelles des années 90. D'un autre côté, les émissions d'actualité diffusées par exemple hier posent également des problèmes. En effet, le contenu de l'actualité ne cesse de changer, et les noms cités, les sujets traités ne sont pas les mêmes. Un mot comme « (Bill) Clinton » est nettement moins cité aujourd'hui qu'en l'an 2000, et le contexte dans lequel il l'est également.

Problématique de la thèse

Un modèle de langage et son vocabulaire qui sont construits à partir de corpus figés, sont pertinents s'ils sont utilisés sur des données figées. À ce propos, les conclusions des campagnes d'évaluation *SDR* sur l'indexation de documents audio [Garofolo *et al.*, 1999], insistent sur la nécessité d'adapter le vocabulaire et le modèle de langage à l'émission à transcrire, notamment en ce qui concerne les différences d'époque qui peuvent exister entre, d'une part, les corpus d'entraînement du modèle de langage, et d'autre part, l'émission à transcrire[7]. L'objectif sous-jacent dans une perspective d'indexation, est que le vocabulaire contienne les mots qui caractérisent de

[7][...] recognition should use adaptation techniques to adjust to changes in the collection language over time.

manière pertinente le document et que la modélisation linguistique permette de les reconnaître[8]. C'est la problématique abordée dans cette thèse.

Les travaux effectués portent sur l'adaptation du vocabulaire et du modèle de langage d'un système de transcription automatique en vue de l'indexation de documents audiovisuels. Le dispositif expérimental mis en place est décrit par le schéma de la figure 1. Envisager l'indexation dans un cadre d'application comme celui de l'INA, définit l'alternative de recherche suivante : soit les documents à transcrire sont plus anciens que les corpus d'entraînement, comme par exemple pour les documents d'archives, soit au contraire, les documents sont plus récents comme cela peut être le cas des émissions recueillies quotidiennement. Dans les deux cas, une des difficultés est de disposer de données au format électronique, qui soient appropriées et contemporaines des documents à transcrire. La solution explorée dans cette thèse est d'utiliser des données provenant de sites internet pour constituer les ressources textuelles pertinentes et contemporaines à l'émission, puis d'appliquer des méthodes d'adaptation du vocabulaire et des modèles de langage utilisant ces ressources.

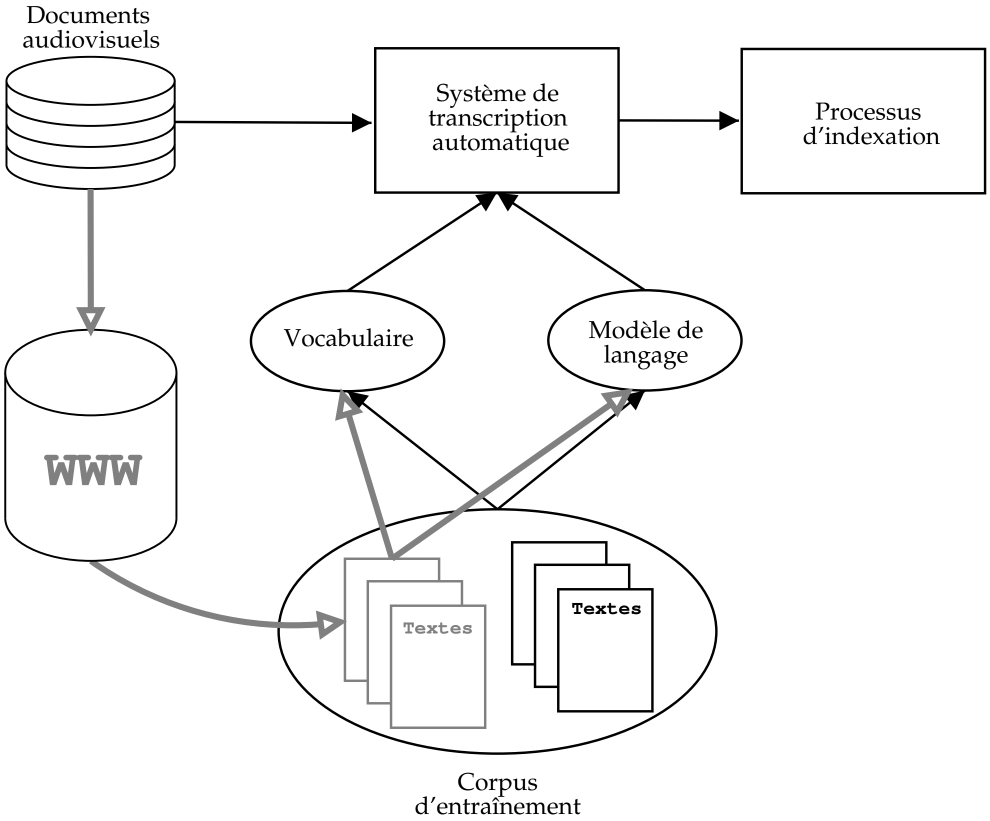


FIG. 1 – Schéma de principe des expérimentations.

[8]Traditional Hub-4-style broadcast news recognizers empoyed only static pre-trained language models. If such a recognizer was used in a real time-longitudinal application, the language in the news and the fixed language model used in the recognizer would diverge, resulting in increasing error rates over time. Such recognizers are incapable of recognizing new words - words likely to be important for retrieval.

Cependant cette solution n'est pas systématiquement envisageable. Soit parce que le corpus à transcrire ne s'y prête pas, soit parce qu'il n'est pas possible de constituer automatiquement des ressources d'adaptation exploitables. Un exemple type est une (ou plusieurs) émission pour laquelle un ensemble de mots qui doit apparaître est connu à l'avance. Mais il n'existe pas de données disponibles permettant de les inclure dans la modélisation linguistique. Pour résoudre ce problème, un modèle de langage à vocabulaire ouvert est développé dans lequel des mots nouveaux peuvent être inclus sans renouveler l'étape de l'entraînement du modèle.

Ce mémoire est divisé en huit chapitres. Le premier présente le point de vue documentaire sur l'indexation audiovisuelle telle qu'elle est pratiquée à l'INA. Cette présentation permet de mieux cerner les objectifs des systèmes d'indexation automatique décrits au deuxième chapitre. Leur description est structurée autour des acquis des principales campagnes d'évaluation organisées par le NIST, et elle met en exergue le rôle fondamental de la transcription automatique de la parole dans les systèmes d'indexation et plus précisément l'impact de son contenu lexical et linguistique. Ces deux premiers chapitres se concluent en définissant la position et les objectifs de mes travaux de thèse à partir des observations sur les pratiques documentaires et de l'étude de l'existant en matière d'indexation automatique audiovisuelle.

Le troisième chapitre décrit le système de transcription automatique de documents audiovisuels du LIMSI, et les fonctions que remplissent ses différentes composantes. Deux d'entre elles (le vocabulaire et le modèle de langage) prépondérantes dans notre problématique, font l'objet des deux chapitres suivants. Le quatrième chapitre aborde les aspects théoriques de la modélisation statistique du langage, et les aspects pratiques de la construction d'un modèle de langage. Il se termine par une étude de l'existant en matière d'adaptation des modèles de langage pour la transcription d'émissions télé et radiodiffusées permettant ainsi de justifier les choix expérimentaux faits par la suite. Puis au chapitre 5, le processus de construction manuelle du vocabulaire est décomposé, afin de proposer un algorithme de construction automatique.

Dans le cadre de la transcription automatique quotidienne d'émissions d'actualité, il est important de pouvoir suivre, au jour le jour, l'évolution lexicale et linguistique de l'actualité. Or les dispositifs et les ressources nécessaires à la construction du vocabulaire et des modèles de langage ne sont pas actuellement compatibles avec de telles exigences. Au sixième chapitre, des algorithmes sont présentés qui permettent d'adapter le vocabulaire à l'actualité en utilisant des textes extraits du Web.

Le chapitre 7 décrit une méthode de construction de modèles de langage qui permet d'introduire des mots nouveaux sans réentraîner les modèles. Un des objectifs de cette méthode est de créer un modèle de langage dont le vocabulaire soit modifiable de manière interactive en fonction de l'émission à transcrire. Enfin, dans le chapitre 8 sont proposées des méthodes d'adaptation des modèles de langage dans les deux cadres d'application choisis lors du positionnement de mes travaux de thèse : pour des documents d'archive et pour la transcription quotidienne d'émissions d'actualité.

Chapitre 1

Pratique documentaire audiovisuelle

L'indexation est indispensable pour faire face à l'inflation d'informations sous toutes ses formes (textes, sons, images). Historiquement dévolue aux bibliothèques et aux documents écrits, elle s'est étendue progressivement aux documents audio et vidéo. Bien que la quantité d'information ne cesse de croître, motivant la recherche sur les systèmes d'indexation automatique, cette pratique reste entièrement manuelle dans beaucoup de cadres d'application. En France, des institutions comme l'Institut National de l'Audiovisuel (INA) ou Radio France, indexent chaque année plus de 100 000 heures de documents audio et vidéo. Elles ont acquis au fil des ans un savoir faire rare en la matière.
Ce chapitre rassemble les observations acquises au cours des rencontres avec les services de documentation qui indexent manuellement les documents audiovisuels. Ces rencontres avaient pour finalité de mieux comprendre le processus d'indexation, et de mieux définir par la suite les objectifs d'un système d'indexation automatique. Ce chapitre débute par une définition de l'indexation. L'ambition n'est pas de définir l'indexation de manière unique, mais d'adopter ici le point de vue documentaire. Puis, les particularités liées à l'objet audiovisuel sont envisagées ainsi que leurs implications. Enfin, le cadre de l'INA est pris comme exemple pour résumer en quoi consiste l'indexation du flux audiovisuel.

1.1 Définition de l'indexation du point de vue de la documentation

L'indexation consiste, en toute généralité, à substituer au document originel une représentation de ce document sous la forme d'une description abrégée, le plus souvent textuelle. Sa finalité est de permettre de repérer rapidement au sein d'un ensemble (ou d'un document), les documents (ou les extraits) pertinents en fonction d'une requête donnée. Cette définition, très générale, est conforme à celle donnée par la norme en vigueur[1] [AFNOR, 1996].

[1] La norme NF Z 47-102 de l'AFNOR (octobre 1993) qui "a pour objet d'établir des méthodologies valables et cohérentes pour la description et la caractérisation des documents à l'aide de représentations de concepts"

Pour les documents analogiques, les données descriptives (ou descripteurs) ont longtemps été gérées indépendamment du support physique auquel elles sont associées. Avec l'utilisation extensive de documents numérisés, les descripteurs et le document coexistent sur un même support. Il devient alors possible d'établir des liens et d'assurer des correspondances entre un des éléments du document (objet, segment, événement) et sa description. Ces données, appelées métadonnées [Hunter and Armstrong, 1999], attachées au document, en facilitent la lecture, la description et l'accès : "l'indexation est une tâche centrale du traitement documentaire puisqu'elle préside à la constitution d'une description qui sert de substitut au contenu dans les phases de recherche et de consultation du document" [Auffret, 2000].

Pour certains, le terme indexation désigne à la fois l'acte de générer cette description abrégée, et les techniques de recherche qui y sont associées. Ceci est particulièrement le cas pour les systèmes d'indexation automatique qui englobent les deux actes en un seul. Bien entendu ces deux phases doivent être pensées en adéquation. Mais dans le monde documentaire, l'usage est de nettement dissocier les deux, indexation et recherche. C'est également le cas dans ce chapitre où seul le processus d'indexation sera abordé.

1.2 Spécificité de l'indexation des documents audiovisuels

La problématique de l'indexation n'est pas nouvelle. Par contre, l'objet audiovisuel est lui récent. Ainsi le processus global de l'indexation est soumis aux spécificités de l'objet audiovisuel : quels sont les index, quelles particularités ont-ils, et à l'autre bout de la chaîne, quelles sont les recherches qui seront effectuées ? Ces questions sont nouvelles car l'usage de documents audiovisuels en tant que produits de grande consommation est récent.

1.2.1 Principales spécificités du document audiovisuel

Les documents audiovisuels présentent un certain nombre de spécificités [Auffret, 2000] par rapport aux autres formes de documents, et notamment aux textes :

- la nature composite. Un document audiovisuel est un document composite, résultant de la production conjointe de l'image, du son, et même parfois d'un texte incrusté. L'interaction de ces différents médias n'est donc pas sans conséquence sur la compréhension de l'ensemble.
- l'icônicité désigne le fait que l'image est une icône de ce qu'elle représente. À la différence du langage, le rapport entre l'icône et la réalité à laquelle elle se réfère est fondé sur un rapport de ressemblance. L'interprétation est arbitraire.
- la polysémie est liée à l'ambiguïté de toute forme d'illustration, elle concerne aussi bien le texte, le discours, l'image fixe ou animée et bien entendu un document audiovisuel. Cette polysémie de l'image est liée à la fois à la temporalité (nécessité de recontextualiser une séquence pour la comprendre), à l'icônicité, et à la nature composite qui multiplie les modalités.

---

indique notamment que "l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse."

La polysémie peut être envisagée principalement à trois moments de la vie d'un document audiovisuel. Lors de la production, les images sont réalisées avec une finalité donnée, liée à un contexte de création. Ensuite, le traitement documentaire induit à la fois la définition d'un usage et un point de vue permettant une certaine lecture du document. Notons que, pour un usage donné, le contexte interprétatif, et donc la lecture, peuvent être variables. Enfin, une séquence audiovisuelle, extraite de son contexte d'origine, peut être réutilisée pour illustrer d'autres propos. Ainsi, un reportage consacré aux conséquences de la guerre du Golfe a montré un cormoran mazouté ; cette image était en fait extraite d'un reportage sur la marée noire, intervenue sur le littoral breton dans les années 70. Le commentaire des images, l'influence de la production du document (du tournage au montage), bref la composition du document audiovisuel, permet de lever les ambiguïtés ou au contraire de biaiser l'interprétation des images.

En résumé, l'indexation des objets dépend du contexte, de ce que l'ensemble du document suggère. De plus, l'information perçue est reformulée au moyen d'un langage écrit, et donc dans une forme différente de celle du document audiovisuel : un document audiovisuel est indexé par du texte.

1.2.2 Une indexation nécessairement subjective

L'indexation des documents audiovisuels est intrinsèquement subjective. Ainsi, dans [Bachimont, 98], indexer c'est réaliser une "paraphrase d'un contenu en une forme sémiotique interprétable permettant de rendre exploitable le contenu indexé dans le cadre d'une pratique donnée (...). L'indexation est donc une interprétation d'un contenu, une réécriture ou reformulation dans une forme propre à l'exploitation dans un contexte."

Par conséquent, du point de vue du traitement documentaire, l'indexation est un processus à la fois subjectif et éphémère qui reflète un usage donné à un instant donné. Pour reprendre les propos de [Veneau, 2002], « d'une part, il ne s'agit pas tant de nier toute possibilité d'indexation que de prendre conscience de sa subjectivité. D'autre part, l'ensemble des subjectivités individuelles est contraint par des pratiques culturelles, des usages plus ou moins figés qui guident l'interprétation ».

1.2.3 Une indexation guidée par l'usage

La recherche et la consultation d'un document peut s'établir à des niveaux d'approches différents selon que l'on s'intéresse à une thématique (l'écologie) ou à un objet spécifique (un oiseau mazouté). Les questions auxquelles nous sommes confrontés sont alors : Quel est le niveau de granularité requis pour décrire un document ? Quel est le point de vue à adopter ? Quelles sont les informations qui fonderont la pertinence de la recherche ou de la consultation du document ? Les réponses sont apportées par l'identification d'un usage prescrit par l'organisme gestionnaire des documents, le domaine d'application visé, etc. Par exemple, une vidéo montrant Henri-Cartier Bresson en Chine au début des années cinquante peut être perçue de multiples façons : l'intérêt peut venir des informations sur le photographe, sur la Chine communiste, ou encore sur les méthodes de filmage utilisées. Il existe non pas un, mais plusieurs scénarios de lecture. Or une description idéale se devrait de prendre en compte tous ces aspects. Puisqu'il n'est pas possible d'épuiser tous les points de vue sur le document, la perspective d'un usage est le seul guide en matière d'indexation.

1.3 Pratique documentaire audiovisuelle : l'INA

Les pratiques de l'indexation et plus généralement la chaîne documentaire audiovisuelle à l'INA ont déjà fait l'objet d'une présentation détaillée dans des travaux précédents [Auffret, 2000, Chap. II]. Quelques éléments utiles sont repris ici, et plus de détails peuvent être trouvé dans le document précédemment cité ainsi que dans [Troncy, 2001].

1.3.1 Missions et quantité

L'Institut National de l'Audiovisuel (INA) a en charge depuis sa création en 1975 la conservation et la gestion des archives professionnelles audiovisuelles et le dépôt légal de l'audiovisuel français. Les archives professionnelles regroupent l'ensemble des émissions produites par le monde audiovisuel français. Ces fonds sont désormais figés et un programme de restauration et de numérisation est en cours. Ce programme de numérisation s'accompagne d'une « reprise d'antériorité » : il s'agit de reprendre la documentation de certaines émissions d'archives dont l'indexation est soit lacunaire soit inexistante. La mission du dépôt légal consiste à capter et indexer le flux diffusé par un ensemble donné de chaînes télévisées et radiodiffusées . Ce flux est souvent désigné par le terme générique de flux entrant. Contrairement aux archives professionnelles, le fonds ne cesse de croître. Une autre mission qui débute pour l'INA est de rendre disponible ses archives. Ainsi "Les fonds conservés au titre de l'archivage professionnel sont communiqués à des fins professionnelles aux producteurs, diffuseurs, éditeurs, établissements éducatifs et culturels. (...) Les fonds conservés au titre du dépôt légal sont communiqués à des fins scientifiques aux étudiants, chercheurs et enseignants"[2]. Dans le domaine de l'indexation, cela se traduit par un programme de « thématisation » décrit au paragraphe 1.3.4.

Mais les missions de l'INA prennent tout leurs sens lorsque les quantités de données manipulées sont évoquées[3]. Les archives professionnelles représentent un fonds de 845 000 heures de télévisions et 974 000 heures de radio, dont environ un million d'heures sont désormais numérisées. La captation du flux entrant représente plus de 100 000 heures par an. Physiquement, tout cela occupe plus d'une centaine de kilomètres de rayonnage.

1.3.2 Documentation d'une émission

Dans le cadre de la captation du flux entrant, l'INA procède à la documentation, c'est-à-dire au catalogage et à l'indexation des documents audiovisuels. Il s'agit pour les documentalistes

[2] Rapport d'activité 2000

[3] Les chiffres cités datent du 31 décembre 2002

d'interpréter le contenu des documents pour en dégager les caractéristiques les plus représentatives, en vue des usages ultérieurs précédemment décrits. L'indexation des documents débouche concrètement sur la rédaction de notices documentaires. Cette notice documentaire est le substitut du document audiovisuel pour les recherches ultérieures.

Segmentation du flux

Le flux entrant est segmenté, selon un processus descendant, de la récupération des grilles de programmes sur une semaine, à l'obtention d'unités de traitement documentaire[4]. Les unités de traitement documentaire sont principalement des émissions ou parfois des tranches horaires.

La plupart du temps, les documentalistes ont la possibilité de visionner le document, sauf dans le cas où les documents sont difficilement manipulables comme par exemple dans le cadre de la reprise d'antériorité. En effet, les anciens supports sont souvent très fragiles, et leurs manipulations et celles des appareils de lectures sont délicates. Les anciennes pratiques, qui avait pour but de limiter au maximum la manipulation des supports, réapparaissent alors : le document a été visionné une fois et un rapport d'écoute (radio) ou de visionnage (télévision) est produit, la suite de la chaîne documentaire n'utilisera plus que ce rapport. Parfois, une documentation écrite est fournie par les diffuseurs[5] ; cependant cela n'est pas systématique, et les informations disponibles sont très variables.

Structuration de l'émission

La documentation d'une émission débute par sa structuration. Il s'agit pour le documentaliste de dégager la structure du document par une segmentation en séquence. Cette structuration permet de sélectionner les séquences qui donneront lieu à une annotation fine et synthétique du contenu. Prenons l'exemple des journaux parlés (émissions d'actualités radiodiffusées indexées par Radio-France) : le service de documentation de l'actualité segmente une émission selon les catégories suivantes :

- un « papier » se définit comme le commentaire ou l'intervention d'un journaliste autre que le présentateur,
- une « annonce » concerne les interventions du présentateur
- un « son » est une interview ou un reportage,
- un « enrobée » ou « pochette surprise » est l'assemblage de papiers et de sons
- et une « image sonore » désigne une illustration sonore comme un bruit de train, foule, ...

Le choix des séquences à annoter est prescrit par les consignes d'indexation. Pour un journal télévisé, la première étape est la segmentation en « plateau/reportage ». Notons que pour un journal qu'il soit télévisé ou parlé, les interventions du présentateur principal ne sont pas annotées.

[4]Pour plus de détails sur ces sujets se référer à [Auffret, 2000].

[5]Dans le cadre de la numérisation, il est prévu de collecter, outre les documents sur support numérisé, un ensemble de données fournies en amont dans la chaîne de production/diffusion (projet Captation).

Indexation d'une séquence audiovisuelle

Six principaux types de traitement, correspondant aux six niveaux de description des documents, ont été rapportés dans [Troncy, 2001]. Ils vont d'informations purement signalétiques concernant la production, la diffusion ou les droits, à des descriptions plus analytiques, du type résumé chronologique et annotations sur « qui parle ? à qui ? de quoi ? où ? quand ? dans quel ordre ? », en passant par des descripteurs de genre définis par un ensemble de termes contrôlés et structurés (le thésaurus).

Enfin, notons que l'INA recueille des documents de types très variés, journaux télévisés, magazines (d'actualités, littéraires, de société, ou sportifs), émissions de variétés, documentaires historiques ou scientifiques, émissions artistiques et musicales. Les documents sont caractérisés selon deux listes d'autorité[6] : la typologie prescrite par l'organisme Médiamétrie[7] utilisée par les publicitaires, et la typologie INA fondée sur les genres, les thèmes et les publics.

1.3.3 De l'analogique vers le numérique

Dans le contexte analogique, la gestion des données descriptives est indépendante de la gestion du support physique. Elles fournissent des clés d'accès au contenu du document. Pour un document audiovisuel analogique, l'accès au support physique doit être rare afin d'en limiter une manipulation lourde et fragilisante. La notice documentaire regroupe l'ensemble des descriptions d'un document. La notice documentaire se substitue au document. Ceci a pour conséquence que ce qui est absent de la notice « n'existe plus ».

Dans le contexte numérique où données de description et document décrit cohabitent sur un même support, il devient possible d'établir des liens et de construire des outils de navigations via les métadonnées. L'accès au document est cette fois-ci illimité et simplifié. Cependant cela ne signifie pas que l'on puisse se contenter d'une indexation rudimentaire de type catalogage sous prétexte que le document devient facilement accessible. Car la contrepartie du passage au numérique est que la quantité de documents accessible ne cesse de croître, et il est toujours primordial de pouvoir sélectionner un document ou un extrait sans avoir à visionner l'ensemble. Le passage de l'analogique vers le numérique transforme la notice documentaire en une interface au document mais l'indexation reste la « partie visible » du document.

1.3.4 Usages de l'indexation

La première utilisation de cette indexation est interne à l'INA. En effet, les activités commerciales de l'INA s'axent désormais sur la thématisation. L'idée est de constituer des "lots" de documents ayant une thématique commune. Par exemple : "Jazz : les grands interprètes ", "La guerre au Proche Orient (le conflit israélo-arabe)". L'objectif est de proposer une offre "clé en main" aux publics mais aussi à terme de réduire le champ d'investigation des documentalistes lors de certaines recherches.

[6]listes de termes normalisées.
[7]www.mediametrie.fr

La deuxième utilisation est de répondre aux requêtes des journalistes, ou des chercheurs qui cherchent des extraits d'émissions. Les requêtes sont alors principalement de deux types : soit la question est de la forme "quelqu'un disant quelque chose, à un moment ou à un endroit donné", soit ce qui est recherché est une illustration audiovisuelle comme par exemple une image du chômage. Le second cas se rapproche de la thématisation et laisse une large initiative aux documentalistes alors que le premier correspond à une requête « simple » devant souvent aboutir à un résultat précis.

Enfin, une troisième utilisation très fréquente consiste à utiliser les fonds de l'INA comme une banque d'images audiovisuelles. Ceci explique la rigueur des descriptions d'image qui qualifient les images par ce qu'elles représentent, mais aussi par les techniques utilisées pour les filmer (cadrage, type de plan, ...). Un exemple fréquent de ce type de requête est l'image d'un président de la République sur le perron de l'Elysée, ou pour reprendre l'exemple du cormoran, un oiseau mazouté.

Conclusion

L'indexation, née du besoin grandissant de recherche d'informations, est une pratique référencée aux usages qui utilise des techniques destinées à représenter (construction d'objets de substitution) et fabriquer des outils de recherche, d'accès et de navigation dans les documents au sein de large corpus. La pratique de l'indexation à l'INA, illustre la possibilité de contraindre la subjectivité de la perception des documents audiovisuels par les usages. Toutefois, il ne s'agit pas d'un problème clos dans la mesure où l'usage et la pratique ne sont pas figés. Ils sont au contraire amenés à évoluer, notamment en fonction des techniques et des outils de production, de stockage, d'indexation et de mise à disposition, liées aux supports numériques. Actuellement, une large part du temps consacré à l'indexation d'un document et consacrée à l'appréhension et à la structuration du document.

L'indexation substitue au document audiovisuel une notice documentaire, à partir de laquelle seront effectuées les recherches. La conséquence fondamentale est que tout ce qui n'est pas dans la notice est en quelque sorte perdue pour l'étape de recherche. A cela s'ajoute le fait que beaucoup d'émission comme les journaux radio ou télévisés ne sont que partiellement annotés. Les requêtes sont principalement des demandes d'illustrations (illustrer le chômage ou la rentrée scolaire par une séquence audiovisuelle), des demandes précises d'images, ou des questions impliquant une personne, un lieu, ou une entreprise qui peuvent se résumer par « ce que dit quelqu'un, quelque part, à telle date, et à tel sujet ». Le niveau des descriptions d'une séquence s'étend donc de la qualification du cadrage aux thèmes abordés en passant par le nom des intervenants. Enfin, l'INA doit faire face dans sa pratique de l'indexation à une grande variété des documents, puisque l'institut traite à la fois le flux entrant en continu, et les documents d'archives qui ne sont pas encore totalement indexées.

Chapitre 2

Indexation automatique de documents audiovisuels

Les recherches en indexation automatique de documents audiovisuels ont été en grande partie fédérées par les campagnes d'évaluation organisées par le NIST (National Institut of Standardization and Normalization) [1]. Plutôt que de se livrer à l'exercice d'un état de l'art des systèmes d'indexation, ce chapitre résume les acquis des principales campagnes d'évaluation, avec comme axe de lecture, les problématiques liées à l'indexation de documents audiovisuels telle qu'elle est pratiquée à l'INA. Mais avant cela, une mise au point sur l'opposition entre indexation automatique et manuelle est nécessaire. La campagne *SDR* (Spoken Document Retrieval) qui initia l'indexation automatique de documents audio est ensuite décrite. Enfin ses extensions récentes à la transcription enrichie et à l'indexation vidéo sont résumées.

2.1 Indexation manuelle *vs* automatique

L'incompréhension est vive entre la communauté scientifique de l'indexation automatique et le monde documentaire. Le chapitre précédent tentait de définir l'indexation du point de vue documentaire et ses implications pratiques. Le passage à la description d'un système d'indexation automatique ne peut se faire sans cette section de transition, qui tente de résumer le point de vue du monde documentaire sur l'indexation automatique.

2.1.1 Automatiser l'indexation manuelle ou indexation automatique

L'indexation manuelle est, pour G. Van Slype [Slype, 1987], une activité fondamentalement distincte à la fois de la classification et de l'indexation automatique, en raison des unités significatives reconnues (des descripteurs au lieu des concepts) et de la sélectivité mise en œuvre lors de l'analyse (une classification synthétique locale au lieu d'une analyse critique globale).

[1] www.nist.gov

Ainsi que le note R.M. Bolle, « idéalement, la vidéo sera automatiquement annotée d'après le résultat de l'interprétation par la machine du contenu sémantique de la vidéo ; toutefois, étant donné l'état de l'art de la vision par ordinateur, une abstraction des données si sophistiquée ne serait pas réaliste dans la pratique. L'ordinateur devrait offrir plutôt une assistance intelligente dans l'annotation manuelle de la vidéo, ou l'ordinateur pourrait réaliser une annotation automatique avec une interprétation sémantique réduite »[Bolle *et al.*, 1998].

En suivant l'idée de G. Van Slype [Slype, 1987], les processus impliqués dans l'indexation automatique et manuelle parviennent presque à s'opposer. En effet, du côté de l'indexation manuelle, les documentalistes sélectionnent parmi les concepts présents (implicitement ou explicitement) dans le document ceux à propos desquels le document apporte une information susceptible d'intéresser les utilisateurs du système documentaire. Une analyse critique, contextuelle et globale du document est donc nécessaire. Inversement, l'analyse automatique souvent fondée sur des méthodes de traitement du signal, renvoie à une information numérique intrinsèquement dépourvue de sens et bien souvent locale. Ainsi, l'indexation manuelle propose une compréhension globale du document permettant une interprétation contextuelle des éléments locaux, tandis que l'analyse numérique des documents voudrait à partir d'indices locaux reconstruire une interprétation globale.

2.1.2 Le passage au sémantique

L'hypothèse sur laquelle se fonde le processus ascendant que se propose de mettre en œuvre l'indexation automatique n'est donc pas triviale. Elle suppose l'existence de liens entre les trois types d'éléments qui constituent un document audiovisuel numérisé : les données sémantiques, les données liées à la forme audiovisuelle, les données physiques (ou numériques). Les données sémantiques sont manipulées sous forme de concepts, objets, segments temporels, etc., par les documentalistes. Les données physiques sont liées au support numérique. Et les données liées à la forme audiovisuelle sont liées à la production et au montage. Dans l'article [Smeulders *et al.*, 2000], des chercheurs sur l'indexation automatique de l'image et la vidéo évoquent un problème « crucial » inhérent au traitement automatique des images, à savoir celui du conflit entre la représentation de l'image et sa signification. Ils définissent ainsi le *sensory gap* : le gap entre un objet dans la réalité et l'information donnée par une description manipulable (dans le sens de données numériques) obtenue à partir d'un enregistrement[2].

Automatique ou semi-automatique, la recherche sur l'indexation assistée par ordinateur s'inscrit dans le cadre de l'indexation manuelle décrite précédemment et doit donc se confronter aux problématiques évoquées, avec la difficulté supplémentaire d'automatiser un traitement *sémantique*, c'est-à-dire d'être capable d'extraire et de gérer, dans la mesure du possible, des informations qui font sens. Cependant, ces critiques s'arrêtent à une vision figée d'un état de l'art en indexation vidéo. Elles ne sont pas suffisantes pour occulter les perspectives en indexation automatique, surtout si l'indexation automatique ne se définit pas comme une indexation manuelle automatisée. Par exemple, les progrès en reconnaissance automatique de la parole ouvrent des

[2]The sensory gap is the gap between the object in the world and the information in a (computional) description derived from a recording of that scene.

perspectives en permettant d'accéder à des informations d'un niveau sémantique plus élevé qu'un histogramme de niveau de gris : les mots.

2.2 Indexation automatique de document audio : la campagne d'évaluation *SDR*

Les recherches en indexation automatique de document audio et audiovisuel ont été guidées par des projet comme Informedia[3], THISL [Abberley *et al.*, 1999], ou OLIVE [de Jong *et al.*, 1999] et surtout par les campagnes d'évaluation américaines organisées dans le cadre des conférences TREC par le NIST. TREC[4] (Texte REtrieval Conference) est autant une conférence qu'une campagne d'évaluation dont la tâche principale est l'indexation et la recherche d'information. La conférence permet l'évaluation de systèmes qui à partir d'une question ou d'une requête, sélectionnent un ensemble de documents. Le cadre correspond bien aux problématiques posées par l'indexation manuelle telles qu'elles ont été présentées précédemment, puisque la collection est connue mais pas les requêtes. La question de la généricité et éventuellement de la pérennité d'une indexation est formellement posée. Les évaluations organisées dans le cadre de TREC se focalisent sur l'accès aux données textuelles dans un contexte multilingue, mais elles s'étendent également à d'autres types de données comme c'est le cas avec *SDR* [5] (Spoken Document Retrieval) puis TRECVID[6] (Vidéo).

2.2.1 Motivation

La motivation des évaluations *SDR* est de permettre le développement de systèmes d'accès à une information non-textuelle telle que l'audio en réunissant les communautés de la reconnaissance automatique de la parole et de la recherche d'information. Les technologies de reconnaissance de la parole permettent de transcrire automatiquement des documents audiovisuels, créant ainsi une interface textuelle avec le signal. La transcription contribue à la constitution d'un document au format texte qui se substitue à l'audio et sur lequel il sera possible d'appliquer des techniques éprouvées d'indexation automatique et de recherche d'information. Le lien avec l'indexation manuelle est ici évident, puisque la démarche préconisée passe par un document de substitution qui peut s'assimiler d'un point de vue fonctionnel à la notice documentaire.

Cette campagne a rencontré un tel succès que d'une part, on peut désormais parler d'un système *SDR*, et d'autre part, les organisateurs ont arrêté cette tâche, jugeant que les systèmes de recherche développés étaient satisfaisants étant données les performances d'une part des systèmes d'indexation, et d'autre part, des systèmes de transcription automatique. Étant donné l'état de l'art en recherche d'information et en RAP, l'accès aux documents audio tel qu'il est défini dans *SDR* , est de leur point de vue un problème résolu. Il leur semble souhaitable de passer à un

[3]http ://www.informedia.cs.cmu.edu
[4]http ://trec.nist.gov/
[5]http ://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm
[6]http ://www-nlpir.nist.gov/projects/trecvid/

niveau thématique avec la tâche *TDT* (*Topic Detection and Tracking*). Cependant, le cadre *SDR* est restreint aux journaux télévisés, et son application aux autres types d'émission n'est pas triviale. De plus, les résultats obtenus sont jugés satisfaisant par le NIST, pour un état de l'art donné des techniques de recherche d'information et de reconnaissance de la parole. Mais ces résultats montrent également qu'un système de type *SDR* ne répond pas de manière totalement satisfaisante à un cadre d'application comme celui décrit au chapitre 1.

2.2.2 Système automatique d'indexation de documents audio

De 1997 à 2000, le cadre d'évaluation a beaucoup évolué cherchant toujours à se rapprocher d'une application réaliste. Les informations données par la suite sont relatives aux deux dernières campagnes *SDR* (TREC-8 en 1999 et TREC-9 en 2000). En pratique, un système *SDR* est la combinaison d'un système de reconnaissance automatique de la parole et d'un système de recherche d'information, comme le montre la figure 2.1. La sortie du système de reconnaissance peut être un treillis de mots (voir au chapitre 3), une liste des meilleures hypothèses, une transcription phonétique, ou la meilleure hypothèse, qui est le cas le plus répandu. L'indexation transforme d'abord le texte en une forme normalisée avec comme étapes, le filtrage à l'aide d'une *stoplist* (liste de termes à retirer, la plupart des termes sont des mots outils ou très fréquents), la lemmatisation, et la normalisation. Puis, pour chaque document un score est calculé. Les scores les plus utilisés sont soit de type Okapi ($TF*IDF$)[7] [Robertson *et al.*, 1994], soit basé sur les probabilités unigrammes. La requête est très souvent un texte court qui ne contient pas nécessairement les index qui se trouvent dans les documents pertinents. L'expansion de requête permet d'augmenter le nombre de termes en effectuant une première requête soit sur le même corpus (*Blind Relevance Feedback*), soit sur un corpus parallèle (*Parallel Blind Relevance Feedback*). Cette première requête permet de sélectionner un premier ensemble de documents. Les index les plus fréquents de ces documents sont sélectionnés pour enrichir la requête.

2.2.3 Spécificité de la parole transcrite

Le système de transcription automatique du LIMSI est décrit au chapitre 3. Pour une description détaillée du système complet présenté par le LIMSI aux évaluations *SDR* de TREC-9, le lecteur peut se reporter aux articles [Gauvain *et al.*, 2000; Gauvain *et al.*, 2001] ou pour le système présenté par l'université de Cambridge à l'article [Johnson *et al.*, 2001].

La sortie d'un système de transcription automatique se distingue des textes « écrits » que les systèmes d'indexation ont l'habitude de traiter. En premier lieu, une transcription contient des erreurs de reconnaissance qui modifient le contenu syntaxique et sémantique du discours. Les effets des erreurs de transcription sont traités au paragraphe 2.2.5. De plus, la parole présente certaines spécificités qui sont dues à la spontanéité du discours, telles que les hésitations, les répétitions, ou les modifications et les corrections du propos. À cela, s'ajoute le fait que la grammaire n'est pas toujours respectée. Enfin, une transcription automatique n'est qu'une suite de mots localisés

[7]Pour un terme donné, TF (*term frequency*) désigne la fréquence du terme dans le document considéré, et IDF (*inverse document frequency*) le logarithme de l'inverse de fréquence des documents contenant ce terme.

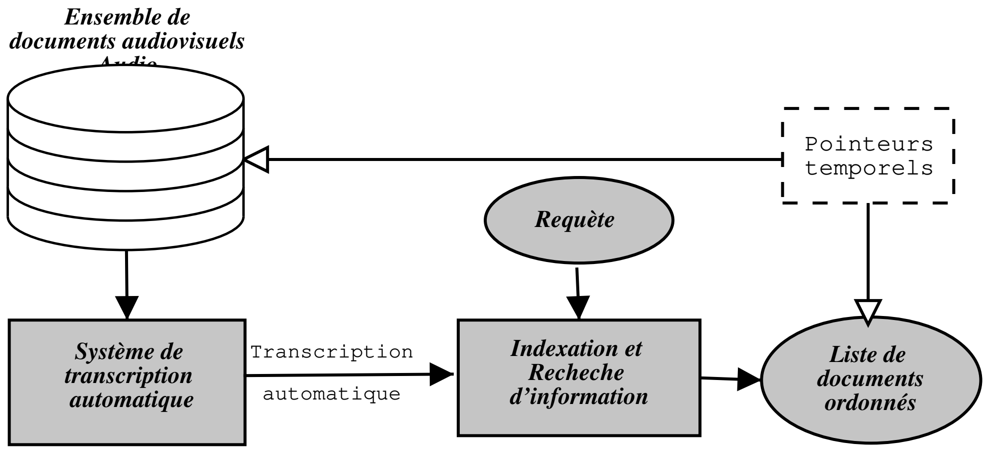


FIG. 2.1 – Architecture type d'un système *SDR*

dans le temps. Cette suite de mots n'est pas segmentée en phrase, ne contient aucune ponctuation et pour certaine langue comme l'anglais la casse est unique ; souvent, le choix d'une casse unique relève de l'habitude et de l'inertie plus que du bon sens. Ces dernières spécificités sont particulièrement dommageables car la ponctuation, la phrase et le paragraphe permettent dans un texte de délimiter les unités thématiques du discours. Et beaucoup de techniques d'extraction d'information s'appuient en partie sur la casse et la ponctuation.

2.2.4 Évaluation d'un système *SDR*

La méthode d'évaluation consiste à demander au système de répondre à une requête par une liste de 1 000 documents ordonnés selon leurs pertinences. La pertinence des réponses est évaluée par des assesseurs humains. Les mesures utilisées sont le rappel et la précision qui se définissent de la manière suivante :

$$\begin{aligned} \text{Rappel} &= \frac{\text{nombre de bonnes réponses du système}}{\text{nombre de bonnes réponses de référence}} \\ \text{Précision} &= \frac{\text{nombre de bonnes réponses du système}}{\text{nombre total de réponses du système}} \end{aligned} \tag{2.1}$$

En faisant varier le nombre de réponses du système, il est possible de tracer l'évolution de la précision en fonction du rappel illustrant ainsi la faculté du système à effectuer un compromis entre donner le plus de bonnes réponses possible (le rappel), et la fiabilité de ces réponses (la précision). Un exemple de courbe est donné à la figure 2.2. L'aire sous la courbe (voir l'aire hachurée à la figure 2.2) caractérise donc le système en une seule grandeur appelée précision moyenne. Cette mesure se nomme *mean average precision* en anglais et est souvent abrégée MAP. Un système idéal a un taux de rappel et une précision qui sont toujours égales à 1 (1 en valeur normalisée ou 100%), indépendement l'un de l'autre. Sur la figure 2.2, cela se traduit par un carré. La valeur optimale de la précision moyenne est l'aire de ce carré, c'est-à-dire 1.

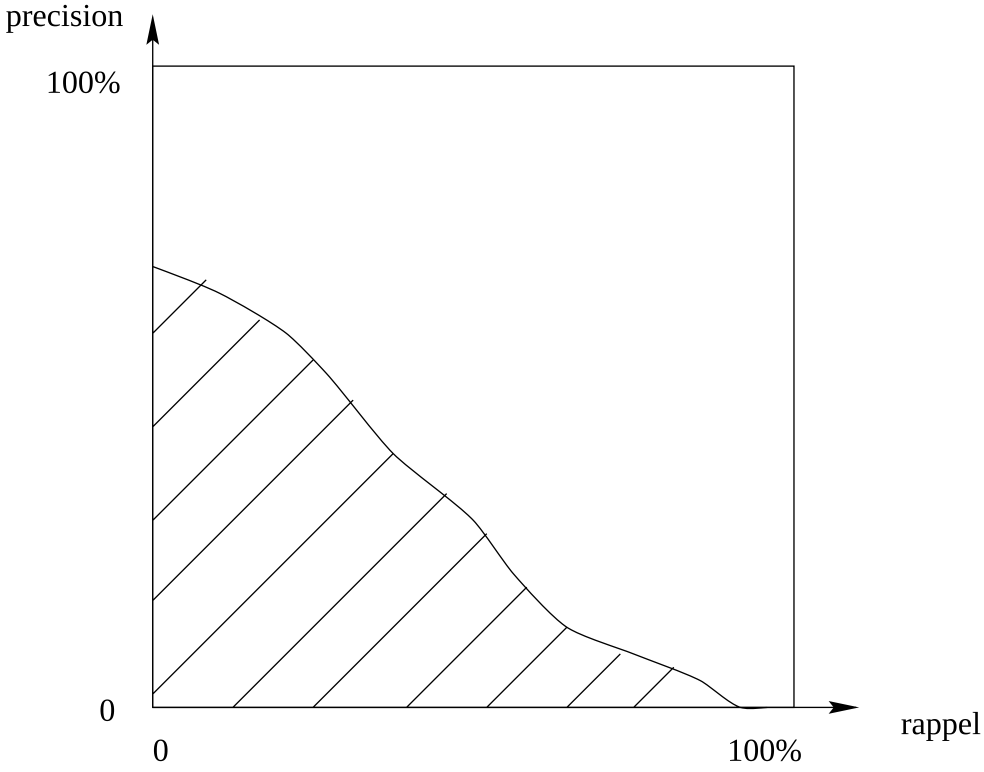


FIG. 2.2 – Exemple de courbe de rappel/précision. La surface hachurée représente la précision moyenne (*mean average precision*).

How and where is nuclear waste stored in New Mexico ?
If we get more income, will we save more or spend more ?
What countries have been accused of human right violations ?
Find reports of fatal air crashes.
What are the latest developments in gun control in the U.S. ?
In particular, what measures are being taken to protect children from guns ?
What percentage of the population is in prison in the U. S. A. and in the E. C. countries ?

TAB. 2.1 – Exemple de requêtes de test utilisées lors des campagnes d'évaluation *SDR*

Le corpus audio contient 557 heures d'émissions d'actualité réparties continûment sur 5 mois de l'année 1998. Ces émissions proviennent de chaînes américaines comme ABC, CNN pour la télévision et de *Voice Of America* ou *Public Radio International* pour la radio. Un corpus parallèle de journaux est également disponible. Un ensemble de 50 requêtes variées ainsi que leurs réponses a été constitué. Le tableau 2.1 en donne quelques exemples.

Puisque la transcription automatique est un flux continu de mots, une réponse est en fait un pointeur temporel qui permet de repérer dans le flux audio un segment qui correspond à la requête. Deux situations ont été envisagées lors de l'évaluation, suivant que les limites des histoires (traduction peu satisfaisante du mot *story*) ou des thèmes sont connues ou pas. Ainsi lorsque les limites sont inconnues, un document réponse est une fenêtre temporelle autour du pointeur. Cette dernière situation correspond à un traitement entièrement automatique des documents audio.

La campagne d'évaluation propose donc un cadre des plus réalistes tant par la définition de la tâche que par la quantité des données, même si ce cadre reste éloigné de celui d'un institut comme l'INA[8]. Cependant, comme les organisateurs l'ont fait remarquer [Garofolo *et al.*, 1999], le cadre est celui d'un traitement sur un corpus figé d'émissions d'actualité. En particulier, les documents audio sont traités en une fois et les recherches sont faites rétrospectivement. Dans le cadre d'un traitement continu de l'actualité sur une longue période, le système de reconnaissance serait amené à subir les changements de thèmes et l'apparition de nouveaux mots (en particulier les noms propres) inhérents à l'actualité. Les auteurs de [Garofolo *et al.*, 1999] prédisent avec peut-être un certain pessimisme un taux d'erreur de reconnaissance qui croit indéfiniment avec le temps. Pour eux, la solution à un problème du type captation et indexation du flux entrant est un système de reconnaissance capable de se mettre à jour, en particulier au niveau du lexique et des modèles de langage utilisés par le système de reconnaissance automatique de la parole [Auzanne *et al.*, 2000].

2.2.5 Effet des erreurs de reconnaissance

Un sous-ensemble de 10 heures du corpus audio a été transcrit manuellement afin d'évaluer les systèmes de reconnaissance. La mesure communément utilisée est le WER (Word Error Rate ou

[8]Les 557 heures du corpus *SDR* sont peu de choses face au millions d'heures des archives audiovisuelles, et les émissions d'actualité sont d'une forme particulière qui ne représente la variété des types d'émission que doit indexer l'INA.

taux d'erreur sur les mots qui est explicité au chapitre 3). Le taux d'erreur du système présenté par le LIMSI à TREC-9 est de 20% [Gauvain *et al.*, 2000].

Lors des évaluations TREC-8 et TREC-9, la corrélation entre le WER et le MAP a été étudiée. Dans l'article [Garofolo *et al.*, 1998], les auteurs observent une certaine corrélation entre ces deux quantités tout en concluant que « le WER est insuffisant pour prédire les performances en recherche d'information »[9]. Intuitivement cela s'explique puisque le WER accorde une importance égale à tous les mots. Or en recherche d'information, un article n'a pas la même importance qu'un nom propre surtout s'il est filtré par une *stoplist*. Les auteurs poursuivent donc en proposant d'autres taux d'erreurs qui prennent en compte l'importance relative des mots. Une erreur de transcription n'a pas le même impact suivant qu'elle porte sur un article indéfini ou sur un nom propre. Ces taux d'erreurs peuvent par exemple être appliqués après une étape de lemmatisation ou après un filtrage de certains mots par une *stoplist*. Mais la mesure ayant le meilleur coefficient de corrélation avec la précision moyenne est le taux d'erreur sur les entités nommées[10] [Garofolo *et al.*, 1998]. Ceci rejoint le paragraphe 1.3.4, puisque que la majorité des requêtes que traite l'INA porte sur des entités nommées.

2.3 Transcription enrichie et indexation vidéo

Les systèmes *SDR* permettent un accès aux données audio par l'intermédiaire de la transcription de la parole. Or, cette transcription n'est pas le seul vecteur d'information dans un document audiovisuel. Dans le cadre de journaux télévisés ou radiodiffusés, la transcription contient une majeure partie de l'information. Mais même dans ce cas favorable, beaucoup d'informations sont perdues. Ainsi aux États Unis avec le projet DARPA EARS[11] ou en France avec le projet ESTER[12], les recherches en indexation de documents audio s'axent sur la transcription enrichie (*Riched Transcript*). Alors que les évaluations organisées dans le cadre des conférences TREC s'étendent à l'indexation des documents vidéo.

2.3.1 Transcription Enrichie

La transcription enrichie participe à la création de métadonnées en essayant d'augmenter les informations liées à la transcription comme par exemple la ponctuation, la désambiguation sémantique, l'annotation sémantique, la reconnaissance, la détection, la caractérisation et le suivi du locuteur.

La problématique de la transcription enrichie peut donc se diviser en trois tâches autour du processus de transcription automatique. La première tâche est bien entendu la transcription avec pour objectif de réduire le taux d'erreur mais aussi d'atteindre le temps réel. Toujours à partir du signal, la deuxième tâche est la segmentation qui se divise en plusieurs catégories. En

[9]This tells us that word error rate alone is insufficient to fully predict retrieval performance

[10]La définition d'une entité nommée est à géométrie variable dans la littérature, ici elle désigne l'ensemble des noms de lieux, de personnes, d'entreprise et d'organisation.

[11]www.darpa.mil/iao/EARS.htm

[12]www.afcp-parole.org/ester

particulier, la segmentation en locuteurs doit permettre de détecter dans un document le nombre de locuteurs et les tours de parole correspondant à chacun. Puis la reconnaissance et le suivi de locuteur permettent d'enrichir la description du document en détectant dans un document un ou plusieurs locuteur(s) connu(s) à l'avance et en déterminant son tour de parole. La dernière tâche est l'extraction d'information. C'est un post-traitement qui utilise les résultats des deux tâches précédentes. Elle reprend des objectifs d'autres campagnes d'évaluation, qui sont appliqués ici à la transcription automatique tels que la ponctuation automatique, la détection des entités nommées, la segmentation en document, la recherche et la détection de thèmes.

2.3.2 Indexation vidéo

Depuis deux ans, TREC et NIST organisent l'évaluation TRECVID[13] qui porte sur la segmentation, l'indexation et la recherche par le contenu de documents vidéo. L'intérêt de cette campagne est double par rapport à un cadre comme celui de l'INA : les corpus vidéo sont des corpus d'archives couvrant une vaste période et un large éventail de type d'émission, et le genre des requêtes recouvre en grande partie les attentes de l'INA en matière d'indexation (comme celles résumées au paragraphe 1.3.4). Pour plus de détails sur les corpus et les thèmes recherchés, le lecteur peut se référer au descriptif de l'évaluation de 2002 [Smeaton *et al.*, 2002], qui est résumée ici.

Trois tâches ont été définies : la localisation et la qualification des changements de plan, l'extraction d'information de l'image, et la recherche d'information. L'évaluation TRECVID est l'équivalent pour les documents audiovisuels des évaluations du type transcription enrichie. La localisation des changements de plan et l'extraction s'apparente aux tâches de segmentation décrites précédemment.

L'extraction d'information a pour objectif la reconnaissance de caractéristiques dites sémantiques de séquences audiovisuelles dont voici quelques exemples :
- scène d'intérieur ou d'extérieur,
- présence d'un visage au moins,
- présence d'une personne au moins,
- paysage de type urbain ou bucolique,
- présence de textes lisibles,
- segment de parole,
- segment contenant des sons instrumentaux,
- segment de monologue.

L'information recherchée est autant dans la bande son que dans les images. Il est intéressant de remarquer que la localisation de telles informations ne donne pas une indexation, mais constitue une première étape. L'ambition d'une telle tâche est de franchir le *sensory gap* évoqué à la section 2.1.2.

La tâche de recherche comporte deux variantes. La première est la recherche manuelle, où un humain rédige une requête, puis les résultats sont évalués. La seconde est la recherche interactive, où une première requête est toujours rédigée par un humain qui affine la requête en fonction des

[13] http ://www-nlpir.nist.gov/projects/trecvid/

résultats. Le critère d'évaluation est la précision moyenne. Le choix des index est libre et les participants peuvent utiliser les résultats des transcriptions automatiques de la parole. NIST a constitué 25 thèmes contenant chacun une description textuelle et parfois un exemple provenant d'un autre média (image ou vidéo principalement). Une requête est une question sur le thème et peut être de la forme : un extrait vidéo ou audio illustrant le thème, une image particulière, ... Les thèmes sont très variés puisque leur intitulé va de « Raymond H. Chandler » à « Un papillon vivant » en passant par « un drapeau américain dans le vent ». Le corpus de test contient environ 40 heures d'émissions de tous types et de toutes dates. Donc la campagne TRECVID propose un cadre d'application qui est plus axé sur l'indexation de documents d'archive alors que pour *SDR*, le type d'émission et l'époque étaient restreints aux émissions d'actualités contemporaines[14]. Pour l'évaluation qui a eu lieu en 2002 (TREC-11), il n'y a pas d'article de synthèse des résultats des différents participants.

Le laboratoire français CLIPS-IMAG a participé à ces évaluations avec, entre autre, l'aide du LIMSI pour la transcription automatique de la parole. Une description complète du système et des résultats se trouve dans l'article [Quénot *et al.*, 2002]. Trois systèmes ont été présentés qui diffèrent par le choix des index. Le premier n'utilise que la transcription automatique, le second que des descripteurs extraits de la vidéo (ceux utilisés dans la tâche extraction d'information) et le troisième est l'union des deux premiers. La conclusion est édifiante : le meilleur système est celui qui utilise uniquement la transcription automatique, le second est l'union des deux premiers systèmes, et très loin derrière se trouve le système qui utilise uniquement les descripteurs extraits de la vidéo. Les informations extraites de la vidéo nuisent donc à l'indexation effectuée sur la transcription automatique. Une conclusion possible est que l'information est prépondérante dans la bande son par rapport à la vidéo. Une autre est que les caractéristiques extraites de l'image sont soit critiquables soit extraites avec encore trop d'erreurs. Dans tous les cas, lorsqu'il s'agit d'indexer automatiquement un document audiovisuel, la transcription de l'audio s'impose comme étant la source d'information la plus fiable et la plus riche. Alors que dans le cadre des campagnes *SDR* la pertinence de la transcription automatique de la parole a été montrée sur un corpus d'émission de type restreint (émissions d'actualité), la campagne TRECVID permet de généraliser cette pertinence à d'autres types d'émissions.

2.4 Conclusion et position de la thèse

Les techniques d'indexation automatique de documents audiovisuels ne semblent pas encore suffisamment avancées pour remplacer, en l'état, l'indexation manuelle. Cependant, les campagnes d'évaluation ont initié plusieurs axes de recherche centrés sur l'extraction d'informations sémantiques du flux audiovisuel, et les problématiques qu'elles abordent rejoignent celle du cadre d'application réel que propose l'INA.

En particulier, les résultats des évaluations TRECVID pour le CLIPS ont montré qu'étant donné les techniques de traitement d'image actuelles, la transcription automatique de la parole constitue l'accès privilégié au flux audiovisuel : le système le plus performant n'utilise que la

[14]En 2003, la campagne d'évaluation TRECVID utilise le même corpus que SDR 99 qui contient des émissions d'actualités contemporaines.

transcription automatique et un moteur d'indexation et de recherche et aucune primitive extraite de l'image. Les campagnes d'évaluation *SDR* ont donc permis l'émergence de systèmes qui constituent toujours l'état de l'art en indexation automatique de documents audiovisuels. Cependant les erreurs de reconnaissance nuisent aux performances d'indexation. Ces erreurs sont, dues d'une part, à l'imperfection des modèles utilisés par le système de transcription, et d'autre part, à la grande variabilité rencontrée dans les documents audiovisuels (variabilité linguistique et acoustique). Ainsi est-il indispensable de développer des méthodes permettant d'adapter le système et ses modèles aux documents à transcrire.

Cette thèse porte sur l'adaptation du vocabulaire et des modèles de langages du système de transcription automatique de documents audio dans le but de les indexer. D'un point de vue de l'indexation automatique, une alternative de recherche fait écho aux missions de l'INA : le traitement des archives *versus* le traitement du flux entrant. Le flux entrant est un « corpus » qui a la particularité de ne pas être figé. En particulier, les thèmes abordés par l'actualité ne cessent de changer et ils sous-tendent les caractéristiques linguistiques et lexicales des émissions qui alimentent ce corpus. Un système de reconnaissance automatique de la parole doit donc être capable de s'adapter à ces évolutions s'il veut être en mesure de transcrire les documents audiovisuels sur une large période et avec une qualité constante. De plus, les mots nouveaux dans l'actualité sont pour la plupart des entités nommées. Et comme l'a conclu le NIST dans le cadre des évaluations *SDR*, les entités nommées conditionnent les performances et la fiabilité d'une indexation fondée sur la transcription automatique.

Le traitement des archives s'inscrit dans ce que l'INA appelle la reprise d'antériorité. Un corpus d'archive a l'avantage d'être connu et fixe. Par contre, comme il couvre une vaste période (par exemple pour l'INA, les documents les plus anciens remontent aux années quarante), il se caractérise par une grande hétérogénéité aux niveaux acoustique, linguistique et lexical. Les modèles statistiques utilisés par le système de reconnaissance automatique de la parole ont été estimés à partir de données d'entraînement récentes qui sont les seules exploitables et disponibles, ceci implique un autre type d'adaptation qui sera abordé dans cette thèse.

L'objectif de cette thèse est donc de développer des méthodes d'adaptation du vocabulaire et des modèles de langage d'un système de transcription de document audiovisuel qu'il soit contemporain ou d'archive. Cette transcription participe aux métadonnées nécessaires à l'indexation automatique du document.

Chapitre 3

Reconnaissance automatique de la parole

Le système de transcription automatique est un des éléments fondamentaux des systèmes d'indexation automatique de documents audiovisuels. Il s'appuie sur une modélisation statistique du processus de génération de la parole. Les recherches dans le domaine sont nombreuses, et le lecteur peut se référer aux ouvrages [de Mori and Bigi, 2002; Gauvain and Lamel, 2003; Jelinek, 2000] pour en avoir une vision complète. Ce chapitre présente le système de transcription automatique de document audiovisuel utilisé dans les travaux décrits par la suite. Les problématiques évoquées ici se restreignent donc à la reconnaissance de la parole continue à grand vocabulaire. Après l'énoncé des principes de la reconnaissance statistique de la parole et la description des modèles utilisés, l'architecture d'un tel système est présentée ainsi que son principal critère d'évaluation.

3.1 Reconnaissance statistique de la parole

La plupart des systèmes reposent sur une modélisation statistique du processus de génération de la parole. Mettre des mots sur un signal équivaut à trouver la séquence de mots $W = w_1, w_2, ..., w_n$ la plus probable étant donné le signal observé. Chaque mot de cette chaîne appartient à l'ensemble fixe et fini que constitue le vocabulaire V. Et en fait de signal, c'est une suite de vecteurs d'observations acoustiques X. Ceci se traduit par l'équation suivante :

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{W}/\boldsymbol{X}) \tag{3.1}$$

où $P(\boldsymbol{W}/\boldsymbol{X})$ est la probabilité de la séquence de mots W étant donné la suite d'observations acoustiques X. L'application de la formule de Bayes permet d'inverser les dépendances et la modélisation « Source/Canal » permet de décomposer le processus de production de la parole comme le montre la figure 3.1. Le message W est généré par un modèle linguistique $\mathrm{P}(\boldsymbol{W})$. Ce message est transformé par le modèle de prononciation $\mathrm{P}(\boldsymbol{H}|\boldsymbol{W})$ en une séquence de phones H.

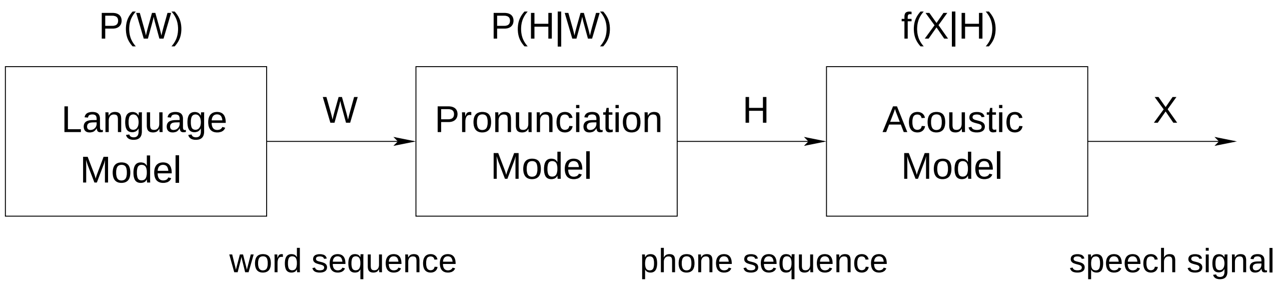


FIG. 3.1 – Modèle source/canal du processus de production de la parole (figure extraite de [Gauvain and Lamel,2003]

Enfin le canal acoustique $\boldsymbol{f}(\boldsymbol{X}|\boldsymbol{H})$ encode H dans le signal $\boldsymbol{X}$. La reconnaissance de la parole consiste alors à trouver la séquence de mots $\boldsymbol{W}^*$ qui maximise la probabilité *a posteriori* de $\boldsymbol{W}$ décomposée de la manière suivante :

$$\hat{W} = \operatorname*{argmax}_{\boldsymbol{W}} \sum_{\boldsymbol{H}} \mathrm{P}(\boldsymbol{W})\mathrm{P}(\boldsymbol{H}|\boldsymbol{W})\boldsymbol{f}(\boldsymbol{X}|\boldsymbol{H}) \tag{3.2}$$

L'équation 3.2 met en évidence les différentes composantes et l'architecture d'un système de reconnaissance automatique de la parole telle qu'elle est représentée à la figure 3.2. Le décodeur cherche la séquence de mots la plus probable en utilisant les modèles acoustiques pour estimer $\boldsymbol{f}(\boldsymbol{X}|\boldsymbol{H})$, un modèle de prononciation pour $\mathrm{P}(\boldsymbol{H}|\boldsymbol{W})$, et un modèle de langage pour la probabilité *a priori* $\mathrm{P}(\boldsymbol{W})$. Donc pour qu'un mot existe du point de vue du décodeur, il est indispensable qu'il fasse partie du vocabulaire, que sa transcription phonétique soit disponible, et que le modèle de langage soit apte à lui attribuer une probabilité.

3.1.1 Codage du signal

Le signal de parole est échantillonné avec une fréquence comprise entre 8 et 16 kHz. Une transformée de Fourier à court terme (algorithme FFT) est appliquée sur une fenêtre d'observation de 30 ms et ce, toutes les 10 ms. Un filtrage est effectué pour mettre le spectre à l'échelle MEL. C'est une échelle perceptive qui modélise à l'aide d'un banc de filtres, la réponse en fréquence du système auditif humain. Deux types de coefficients peuvent alors être calculés : les coefficients MFCC [Davis and Mermelstein, 1980] sont issus d'une transformation de Fourier inverse appliquée au logarithme du spectre de puissance, alors que les coefficients PLP sont obtenus par transformation de Fourier inverse de la racine cubique du spectre de puissance suivie d'une analyse par prédiction linéaire [Hermansky, 1990]. Généralement, les douze premiers coefficients sont retenus, auxquels s'ajoute le logarithme de l'énergie normalisée. Treize coefficients sont ainsi obtenus qui représentent un intervalle de signal de 10 ms. La dimension du vecteur acoustique est finalement augmentée à 39 composantes en ajoutant l'approximation de la dérivée première et seconde des treize coefficients (les coefficients delta et delta-delta).

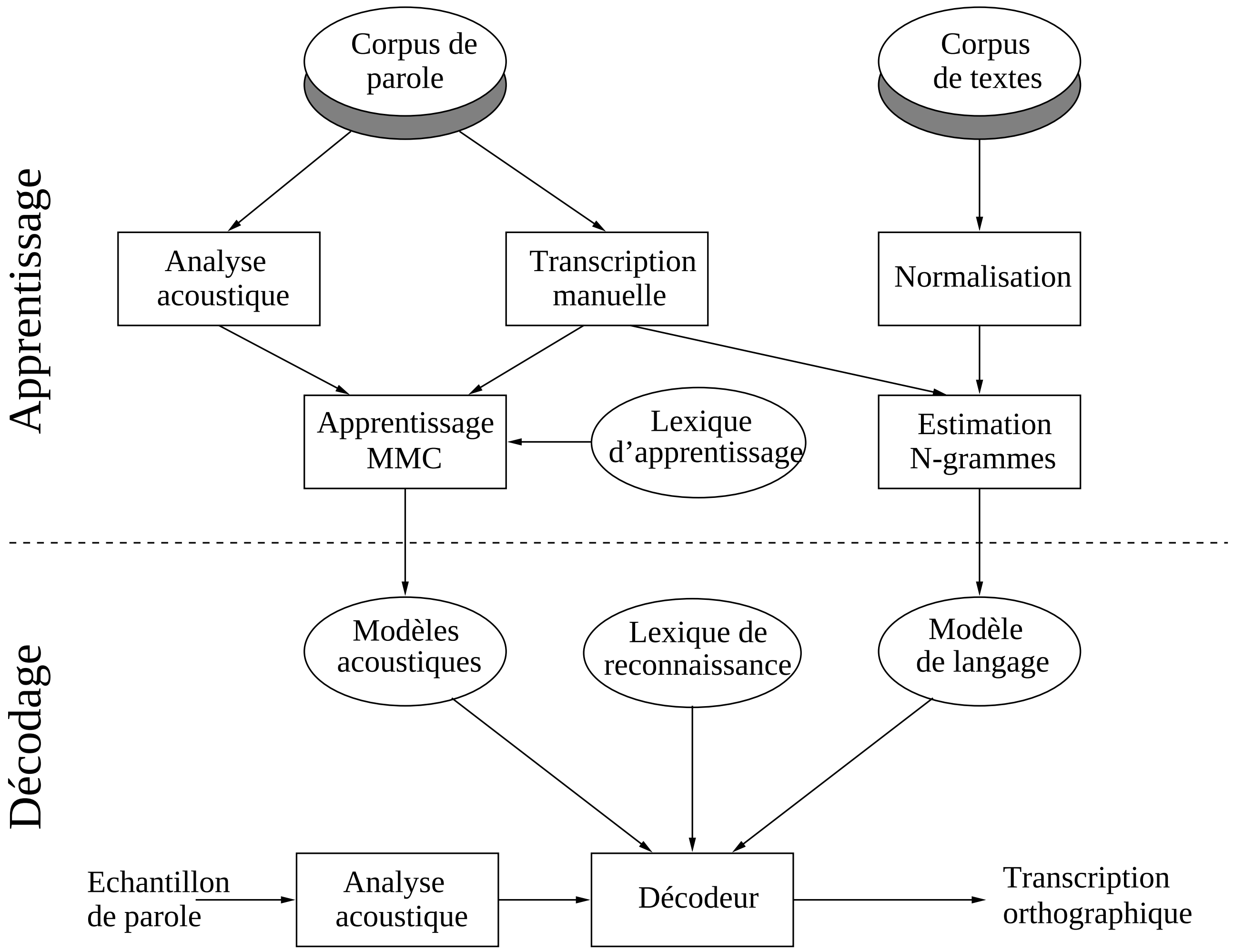


FIG. 3.2 – Architecture d'un système de reconnaissance automatique de la parole (figure extraite de [Gauvain and Lamel,2002]

3.1.2 Modélisation acoustique

Les modèles de Markov Cachés (MMC) sont utilisés dans la plupart des systèmes de reconnaissance automatique de la parole pour modéliser les séquences de vecteurs acoustiques à partir d'unités acoustiques. L'unité acoustique élémentaire est l'allophone : un phonème en contexte modélisé par un MMC à 3 états tel qu'il est représenté à la figure 3.3. À chaque état est associée une densité de probabilité sur les vecteurs acoustiques. Le processus stochastique de génération d'une suite de T vecteurs acoustiques $X = (\mathbf{x}_1, ..., \mathbf{x}_T)$ par la séquence d'états $\mathbf{s} = (s_0, ..., s_T)$ est alors décrit par la densité de probabilité conjointe $f(X, \mathbf{s})$:

$$f(X, \mathbf{s}) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} f(\mathbf{x}_t | s_t) \tag{3.3}$$

où π_s est la probabilité initiale de l'état s, a_{ij} est la probabilité de transition de l'état i à l'état j, et $f(\cdot|s)$ est la densité d'émission associée à l'état s. L'entraînement des modèles acoustiques consiste à estimer pour chaque MMC les probabilités de transition et les densités d'émission associées à chacun de ses états. Pour ces dernières, le choix se porte généralement sur une estimation paramétrique qui fait l'approximation d'un mélange de gaussiennes. La densité d'émission s'écrit alors :

$$f(\mathbf{x}_t | s) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}_t | \mathbf{m}_{sk}, \Sigma_{sk}) \tag{3.4}$$

où ω_k est le poids associé à la gaussienne k avec un vecteur moyen m_k et une matrice de covariance Σ_k (souvent supposée diagonale). La méthode d'entraînement est le plus souvent fondée sur le critère du maximum de vraisemblance qui consiste à estimer les valeurs des paramètres qui maximisent la vraisemblance des données d'apprentissage. L'algorithme EM (Expectation Maximization) [Dempster *et al.*, 1977] est couramment utilisé pour effectuer cette estimation. Les états du modèle sont alignés sur les données d'apprentissage en affectant chaque vecteur à un état grâce à l'algorithme de Viterbi. Et les valeurs des paramètres sont obtenues via les formules de réestimation de Baum-Welsh [Baum *et al.*, 1970].

Le critère du maximum de vraisemblance conduit à l'estimation de modèles représentatifs des données d'apprentissage et des locuteurs présents dans ces données. Cependant il existe des choix et des méthodes d'adaptation qui permettent aux modèles de s'affranchir en partie des données d'apprentissage [Gauvain and Lee, 1994; Leggetter and Woodland, 1995]. En particulier, en ce qui concerne l'indépendance vis-à-vis du locuteur, les corpus d'entraînement contiennent les énoncés d'au moins une centaine de locuteurs différents. De plus de meilleurs résultats sont obtenus en modélisant séparément les locutrices et les locuteurs.

3.1.3 Lexique de reconnaissance

Le lexique de reconnaissance (également appelé dictionnaire phonétique, dictionnaire de prononciation, lexique phonétisé ou simplement lexique) est le lien entre la modélisation linguistique et la modélisation acoustique. Il regroupe tous les mots du vocabulaire et leurs transcriptions

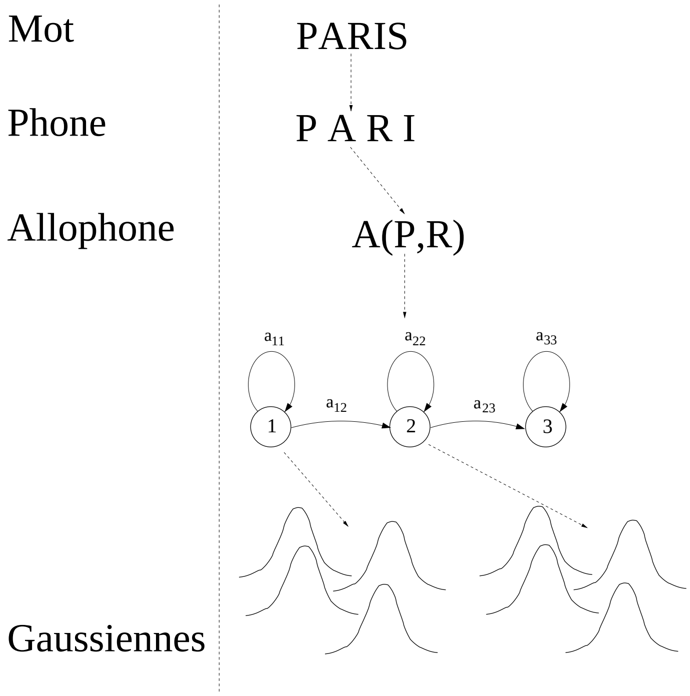


FIG. 3.3 – Modélisation d'un mot : un mot est transcrit phonétiquement, puis chaque phonème est converti en allophone et à chaque allophone correspond à un MMC (figure extraite de [Gauvain and Lamel,2002]

phonétiques. Il suppose donc le choix préalable des entrées lexicales et du jeu de phonèmes utilisé pour les décrire. Le choix du jeu de phonèmes est lié à la modélisation acoustique et à la langue (par exemple nous utilisons environ 45 phonèmes pour l'anglais, 49 pour l'allemand, 33 pour le français, et seulement 26 pour l'espagnol). Le choix du vocabulaire est lié à la modélisation linguistique et sera plus amplement décrit aux chapitres 5 et 4. Les dictionnaires de prononciation sont généralement construits manuellement par des experts. La figure 3.3 représente la chaîne de modélisation d'un mot du vocabulaire : un mot est transcrit phonétiquement, puis chaque phonème est converti en allophone et chaque allophone correspond à un MMC. En pratique, chaque entrée lexicale est décrite à l'aide d'une séquence d'éléments choisis parmi 33 phonèmes, auxquels s'ajoutent 3 symboles spécifiques pour représenter les hésitations, les respirations et les silences.

3.1.4 Modélisation linguistique

Un modèle de langage tente de capturer les contraintes sémantiques et syntaxiques du langage via l'étude de ses régularités. La finalité est d'estimer la probabilité *a priori* $\mathrm{P}(\boldsymbol{W})$ de toute séquence de mots extraites du vocabulaire. La construction d'un modèle de langage est plus amplement détaillée au chapitre 4 qui lui est consacré. La plupart des systèmes utilisent des modèles *n-gram* de mots qui prédisent un mot en fonction des $n-1$ précédents. En général n

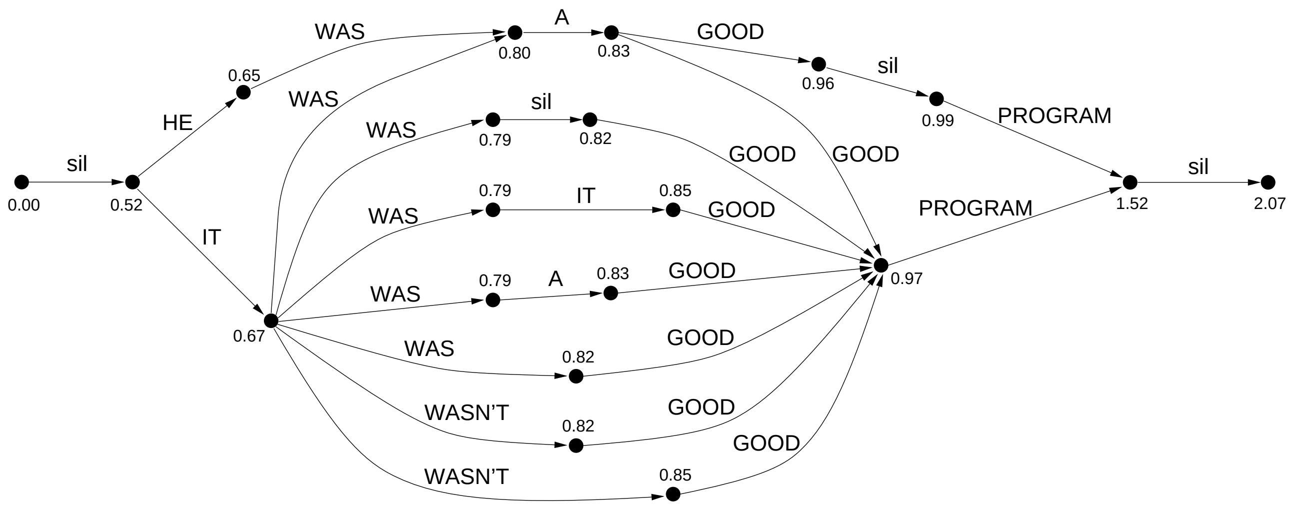


FIG. 3.4 – Exemple de treillis de mots. Les scores linguistiques et acoustiques ne sont pas représentés (figure extraite de [Gauvain and Lamel,2002]

varie de 2 à 4, donnant un modèle nommé respectivement bigramme, trigramme et quadrigramme.

3.1.5 Décodeur

Le décodeur a la lourde tâche de déterminer la suite de mots ayant la probabilité la plus élevée à partir d'une séquence de vecteurs acoustiques, étant donné le lexique, les modèles acoustiques et linguistiques. Pour cela, il envisage un certain nombre d'hypothèses parmi lesquelles, il choisit la plus probable. Dans le cadre d'un système de reconnaissance automatique de la parole à grand vocabulaire (typiquement 65 000 mots), il est évidemment impossible d'envisager toutes les hypothèses. La plupart des décodeurs fonctionnent en plusieurs passes ce qui permet de réduire le temps de calcul et les besoins en mémoire en introduisant les connaissances progressivement.

D'une passe à l'autre, l'espace de recherche se transmet via un treillis de mots[1]. Un treillis est un graphe acyclique dirigé (*Direct Acyclic Graph*) où chaque nœud correspond à un instant donné, et chaque arc est une hypothèse de mot pondérée par un score linguistique et un score acoustique. Ces scores sont estimés à partir des modèles linguistiques et acoustiques. Chaque segment de parole à décoder est donc représenté selon un treillis de mots comme le montre figure 3.4. Un treillis de mots est une représentation compacte des hypothèses les plus probables qui permet un décodage rapide en plusieurs passes, l'estimation de mesures de confiance et l'application d'autres sources de connaissances. Ainsi il devient très simple de modifier par exemple le modèle de langage en remplaçant simplement les scores linguistiques.

Le décodeur utilisé par la suite procède en trois passes. Tout d'abord un premier treillis de mots est généré avec un modèle de langage trigramme. Ce treillis est utilisé pour adapter

[1]Le terme treillis est généralement utilisé comme traduction du mot anglais *lattice*. Cependant ce terme est également utilisé dans d'autres disciplines, comme en mathématique, avec un sens différent.

les modèles acoustiques. La deuxième passe redéveloppe un nouveau treillis avec les modèles acoustiques adaptés et un modèle de langage bigramme. Enfin, en guise de troisième passe, le treillis est recalculé avec un modèle de langage quadrigramme. Pour les trois passes, trois modèles de langages sont utilisés. Il s'agit sauf spécification contraire, du même modèle de langage quadrigramme, mais restreint selon les besoins aux modèles trigramme et bigramme.

3.2 Système de transcription automatique

La section précédente décrit succinctement comment un système de reconnaissance automatique de la parole décode un segment de parole. Or une émission télé-radiodiffusée est un flux de données qui n'est pas segmenté et qui ne contient pas uniquement de la parole. Le premier élément d'un système de transcription automatique de documents télé-radiodiffusés est le « partitionneur » dont le rôle est justement de préparer le flux de données pour l'étape de reconnaissance de la parole. Cette section décrit dans un premier temps le partitionneur et dans un second temps, le principal critère d'évaluation d'un système de transcription automatique.

3.2.1 Le partitionneur

Le *partitionneur* effectue la segmentation du flux audio en détectant les changements de conditions acoustiques. Chaque segment est classé selon trois catégories : parole, musique, et bruit. Les segments de parole sont ensuite étiquetés en locuteur, bande passante et genre du locuteur. L'étiquetage en locuteur utilise une procédure itérative qui regroupe les segments par locuteur. Le résultat est un ensemble de segments regroupés suivant leurs attributs acoustiques, ce qui permettra par la suite au système de reconnaissance automatique de la parole de choisir les modèles acoustiques les plus appropriés pour décoder chaque segment et d'adapter les modèles entre deux passes.

3.2.2 Taux d'erreur sur les mots

L'évaluation d'un système de reconnaissance automatique de la parole consiste en la comparaison entre la transcription de référence et celle fournie par le système. Trois types d'erreur sont pris en compte. Pour cela, les deux transcriptions sont alignées entre elles et chaque mot de l'une a son correspondant dans l'autre. Lorsque deux correspondants ne sont pas les mêmes, il y a une erreur dite de substitution. Si dans la transcription de référence se trouve un mot n'ayant pas de correspondant dans la transcription automatique, alors cela signifie que le système a oublié ce mot. C'est une élision. Enfin l'erreur symétrique est une insertion.

Le taux d'erreur sur les mots est la mesure la plus employée pour évaluer un système de reconnaissance automatique de la parole que ce soit dans le cadre de la dictée vocale ou de la transcription automatique de document audio. Ce taux se calcule à partir du compte des différentes erreurs comme il suit :

$$\text{Taux d'erreur sur les mots} = \frac{\#\text{substitution}+\#\text{insertion}+\#\text{élision}}{\#\text{ mots dans la référence}} \tag{3.5}$$

Réf :	U. N. SECRETARY GENERAL	KOFI ANNAN IS	EXPECTED TO MEET
Hyp :	U. N. SECRETARY GENERAL	KOBE ON AND THAT'S	EXPECTED TO MEET
Réf :	WITH MORE IRAQI LEADERS	THIS WEEKEND FRIDAY	HE SPOKE WITH IRAQI
Hyp :	WITH MORE IRAQI LEADERS	THIS WEEKEND FRIDAY	HE SPOKE WITH IRAQI
Réf :	DEPUTY PRIME MINISTER	TARIQ AZIZ	
Hyp :	DEPUTY PRESIDENT	TARIQ AZIZ	

TAB. 3.1 – Exemple de transcription d'émission d'information américaine (CNN Headline) extraite du corpus SDR 99. La ligne « Réf » désigne la transcription de référence, et « Hyp » l'hypothèse du système de reconnaissance.

Le taux d'erreur sur les mots est très couramment abrégé WER (*Word Error Rate*). Cette métrique présente certains inconvénients. Le premier est son coût puisqu'il faut pour calculer ce taux , transcrire manuellement et finement l'émission. De plus, les comptes utilisés sont indépendants des mots sur lesquels portent les erreurs. Cette mesure considère donc que les erreurs sont toutes préjudiciables de la même manière, et cela indépendemment du cadre d'application. Cet aspect a déjà été brièvement évoqué au paragraphe 2.2.5 dans le cadre de l'indexation automatique.

Conclusion

Dans ce chapitre, le système de transcription automatique de document audio est présenté. Ce système est celui qui sera utilisé dans les expériences de reconnaissances décrites aux chapitres 7 et 8. Ce système est également équivalent à ceux utilisés lors des campagne d'évaluation *SDR* ou TRECVID (voir chapitre 2). Un système à l'état de l'art pour de la transcription automatique de documents d'actualité américains comme celui du LIMSI a un taux d'erreur de l'ordre de 10%. Pour le français, il n'y pas pas de campagne d'évaluation récente, cependant le taux d'erreur du système du LIMSI peut être estimé à environ 20%. Cette différence de performance n'est pas liée à la langue, mais à la disponibilité de données d'entraînement dont les quantités ne sont pas les mêmes pour l'anglais et pour le français.

Les tableaux 3.1 et 3.2 montrent des exemples de transcriptions automatiques pour une émission d'actualité américaine et un document d'archive de guerre française. Les erreurs de transcription illustrent certaines faiblesses de la modélisation linguistique. Concernant l'extrait de CNN, le modèle de langage n'est pas adapté à l'actualité (si « TARIQ AZIZ » est bien reconnu, sa fonction par contre est erronée). Pour la transcription du document d'archive, d'une part, le vocabulaire du modèle de langage n'est pas adapté à l'époque du document (Disney Land n'existait pas encore dans les années 40), et d'autre part, beaucoup des erreurs sont dues aux conditions acoustiques difficiles de l'enregistrement.

Réf :	mais avant	même d' avoir reçu la réponse de l' amiral français
Hyp :	des ardents	même d' avoir reçu à bord de la mer rappelle
Réf :	les Anglais ouvrent les hostilités en faisant	miner l' entrée de la rade par leurs avions
Hyp :	les Anglais ouvrent les hostilités en faisant	Disneyland et de la vapeur d' action

TAB. 3.2 – Exemple de transcription d'émission d'archive

Chapitre 4

Modélisation statistique du langage

Un modèle de langage (ML) statistique tel qu'il est utilisé en reconnaissance automatique de la parole est une distribution de probabilité sur l'ensemble des phrases possibles. Son rôle est de capturer les régularités du langage naturel au moyen de techniques d'estimation statistique appliquées sur des données d'entraînement, à savoir du texte. Être capable d'assigner une probabilité à une séquence de mots est indispensable dans diverses applications, comme par exemple, la reconnaissance de la parole, pour laquelle furent réalisées les premières modélisations, mais aussi la traduction automatique [Brown *et al.*, 1990] ou la classification de documents [Ponte and Croft, 1998]. L'enjeu est donc important et les recherches sont nombreuses sur le sujet. Malgré cela, le modèle « historique » *n-gram* de mots , introduit en 1976 [Jelinek, 1976], constitue toujours l'état de l'art. Pourtant, ce modèle des plus simples, ne prend pas en compte le fait qu'il modélise le langage naturel, et que le langage naturel est structuré. Les tentatives visant justement à introduire des connaissances liées au langage naturel ont permis d'obtenir jusqu'à présent des performances qui sont au mieux comparables à celles d'un modèle *n-gram* entraîné sur des corpus de centaines de millions de mots.

Aborder la modélisation du langage dans le cadre de la transcription automatique de documents audiovisuels peut donc conduire à un état de l'art très court. En effet, l'observation des systèmes de transcription les plus performants montre qu'il est possible de se borner à l'étude du modèle *n-gram* de mots et de classes. Mais l'efficacité de ces modèles dépend de la quantité de données d'entraînement disponibles et de leur utilisation. Le processus de construction se révèle donc lourd et coûteux. Cette lourdeur et ce coût motivent l'étude des méthodes d'adaptation qui tentent de maintenir les performances des modèles sans pour autant entamer à nouveau une construction laborieuse.

La première section introduit le cadre théorique de la modélisation statistique du langage et les différents types de modèles utilisés pour la reconnaissance automatique de la parole dans les documents audiovisuels. Les deux principales techniques de combinaison de modèles et d'introduction de contraintes sont ensuite présentées à la section 4.2, et les critères d'évaluation des modèles de langage sont définis à la section 4.3. Puis la section 4.4 décrit le processus de construction du modèle de langage de référence. Enfin la section 4.5 présente les problématiques et les solutions existantes liées à l'adaptation des modèles de langage.

4.1 Modélisation statistique du langage

Pour la reconnaissance de la parole, le rôle d'un modèle de langage est d'estimer la probabilité *a priori* $P(\boldsymbol{W})$ de la séquence de n mots $\boldsymbol{W}$, comme le décrit l'équation 3.2. Les mots constituant cette chaîne appartiennent à un vocabulaire fini :

$$\boldsymbol{W} = (w_1, w_2, ..., w_n) \text{ , avec } \; w_i \in V \tag{4.1}$$

Grâce à la règle de Bayes, cette probabilité s'écrit :

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} P(w_i | w_1, w_2, ... w_{i-1}) \tag{4.2}$$

La suite de mots $(w_1, w_2, ... w_{i-1})$ est l'historique h_i du i-ème mot w_i de la séquence $\boldsymbol{W}$, alors que par la suite, w désignera un mot à valeur dans le vocabulaire et h l'historique qui lui est associé. Dans le cadre d'une modélisation « Source/Canal » de production de la parole telle qu'elle est présentée à la section 3.1, l'équation 4.2 signifie que l'hypothèse est faite d'un locuteur qui détermine le i-*ème* mot de sa phrase en fonction de l'historique complet de ce qu'il vient de dire. Il peut être utile et réaliste de faire l'approximation consistant à conditionner le mot w_i par une classe d'équivalence de l'historique $\Phi(w_1, w_2, ... w_{i-1})$

$$P(\boldsymbol{W}) \approx \prod_{i=1}^{n} P(w_i | \Phi(w_1, w_2, ... w_{i-1})) \tag{4.3}$$

Ce formalisme est plus amplement détaillé dans l'ouvrage de F. Jelinek [Jelinek, 2000]. Pour reprendre d'ailleurs ses propos, tout l'art de la modélisation du langage consiste à déterminer cette fonction d'équivalence Φ et une méthode pour estimer les probabilités associées[1]. De manière générale, un historique désigne la séquence qui précède le mot considéré et est défini par la fonction Φ.

Chaque probabilité $P(w_i | \Phi(w_1, w_2, ... w_{i-1}))$ est estimée en général à partir d'observations faites sur des corpus d'entraînement. Ce processus est plus amplement abordé au paragraphe 4.4.3. Cependant, l'utilisation de classes d'équivalence se révèle être indispensable pour des raisons d'observabilité. Prenons l'exemple d'un modèle de langage construit sur un vocabulaire de $N =$ 65 000, chaque mot w_i a alors potentiellement N prédécesseurs possibles. L'ensemble des séquences possibles de i mots est de N^i. Donc pour une séquence de deux mots, l'historique est réduit à un mot (un bigramme), cela fait $65\ 000^2 = 4\ 225\ 000\ 000$ possibilités qui sont autant de probabilités à estimer et donc de séquences de mots à observer. Pour une séquence de trois mots ($i = 3$), le nombre de possibilités devient environ $65\ 000^3 = 2,74 \times 10^{14}$. En dehors même du fait qu'il est difficile de manipuler un modèle de langage d'une taille aussi importante, estimer les probabilités de manière fiable est un défi. En effet, même avec des corpus d'entraînement de l'ordre de la centaine de millions de mots, beaucoup d'historiques possibles ne seront peu ou pas observés. La fonction Φ est donc indispensable pour regrouper les historiques en classes d'équivalence réduisant ainsi le nombre d'observations nécessaire, et pour permettre une estimation fiable des probabilités.

[1]The art of language modeling consists of determining the appropriate equivalence classification Φ and a method of estimating the probabilities $P(w_i | \Phi(w_1, w_2, ... w_{i-1}))$.

4.1.1 Le modèle *n-gram* de mots

Le modèle *n-gram* est le plus utilisé en reconnaissance de la parole : c'est le modèle « historique », le plus simple, et ses performances constituent encore l'état de l'art. Le modèle *n-gram* réduit la dimension du problème de l'estimation, en modélisant le langage par une source markovienne d'ordre $n-1$:

$$P(w_i|w_1, w_2, ...w_{i-1}) \approx P(w_i|w_{i-n+1}, ...w_{i-1}) \tag{4.4}$$

Le choix de la fonction d'equivalence Φ est que les historiques se terminant par les mêmes $n-1$ mots sont équivalents. L'estimation des probabilités continue de poser problème, ne serait-ce que pour un modèle bigramme ou trigramme, même avec des corpus d'entraînement dépassant la centaine de millions de mots. Le langage naturel respecte la loi de Zipf[2] [Manning and Schütze, 1999] : dans un texte, si les mots sont classés selon leurs fréquences décroissantes, le rang r d'un mot est inversement proportionnel à fréquence f :

$$f \approx \frac{K}{r}, \tag{4.5}$$

où K est une constante de proportionnalité. La figure 4.1 illustre la loi de Zipf et représente la fréquence en fonction du rang lexical observé sur deux corpus différents. Le premier regroupe des textes de journaux et le second des transcriptions d'émissions télé et radiodiffusées. Ces deux corpus sont décrits plus en détail au paragraphe 4.4.1 et plus particulièrement par le tableau 4.5. Pour les deux corpus, les courbes sont similaires et l'apparente translation entre les deux est due à la différence de taille entre les deux corpus.

Ces deux courbes illustrent le fait qu'à la constante K près, le deuxième mot a donc une fréquence deux fois moins importante que le mot de rang 1 et que le mot de rang 100 a une fréquence 10 fois plus élevée que la fréquence classée au rang 1000. George Kingsley Zipf y voit la manifestation de ce qu'il nomme le principe de moindre effort, qui selon lui caractérise toutes les activités humaines [3]. Ainsi ce principe se généralise aux n-grammes pour $n > 1$, et la figure 4.2 le montre d'une manière différente en représentant les « fréquences des fréquences » des quadrigrammes observés sur les deux corpus (journaux et transcriptions). Les deux courbes sont là aussi similaires et toutes les deux peuvent se décomposer en deux parties. La première partie est la partie linéaire, elle se situe au niveau des fréquences faibles (environ de 1 à 100 pour les journaux). Cette partie montre que la fréquence la plus représentée pour les quadrigrammes est la fréquence unité. Puis plus la fréquence augmente, moins elle est représentée. La seconde partie (les fréquences supérieures à 100 pour les journaux) ne concerne que 0,1% des quadrigrammes. La figure 4.3 est un agrandissement de la figure 4.2 pour les observations faites sur les journaux. Cette figure permet d'observer que presque toutes les hautes fréquences ne concernent qu'un quadrigramme (éventuellement deux pour quelques cas). De plus, dans les hautes fréquences, il est visible que peu de fréquences sont utilisées puisque sur les 90 000 valeurs possibles entre 10 000 et 100 000, seules 290 (0,3%) sont observées.

[2]Le site http ://linkage.rockefeller.edu/wli/zipf/ est dédié à la loi de Zipf.

[3]En effet, un locuteur minimise son effort pour s'exprimer en utilisant la plupart du temps que très peu de mots.

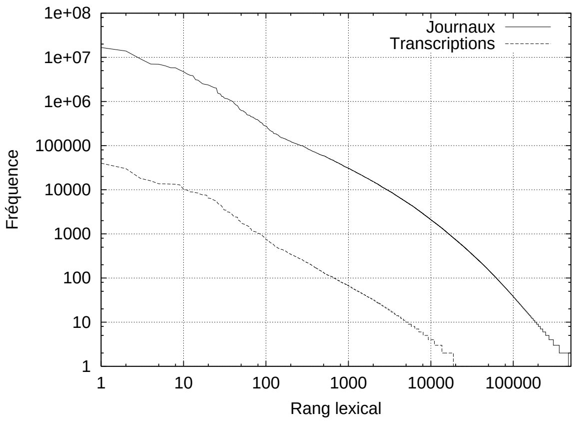


FIG. 4.1 – Observation de la loi de Zipf sur deux corpus : un corpus de textes de journaux et un corpus de transcriptions d'émissions télé et radiodiffusées. Les échelles sont logarithmiques.

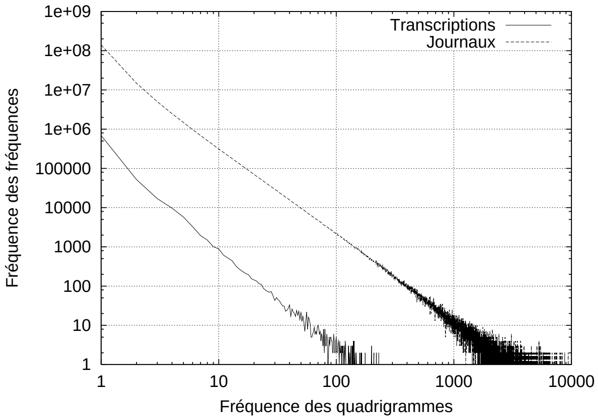


FIG. 4.2 – Fréquences des fréquences des quadrigrammes aobservés sur les corpus de journaux et de transcriptions. Les échelles sont logarithmiques.

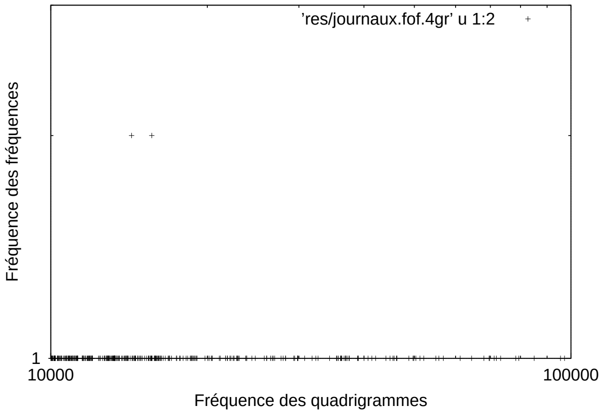


FIG. 4.3 – Fréquences des fréquences des quadrigrammes observés sur le corpus de journaux. Les échelles sont logarithmiques.

La loi de Zipf a pour conséquence que la plupart des n-grammes apparaissent une fois et que les autres ont des nombres d'occurences faibles et équivalents. Pour pallier ce problème, différentes techniques de lissage (ou de *smoothing*) ont été développées. La construction d'un modèle *n-gram* de mots est plus amplement traitée au paragraphe 4.4.

Mais au-delà des problèmes d'estimations des probabilités, la simplicité d'un modèle *n-gram* semble très éloignée de la complexité du langage naturel et il est intéressant de préciser en quoi. Dès le départ, l'hypothèse fondatrice d'indépendance est fausse : si un mot peut être prédit uniquement à partir des $n-1$ précédents, c'est qu'il est indépendant des autres. Or pour une phrase, cette hypothèse n'est pas vérifiée si elle dépasse les n mots. En dehors de cette fenêtre de $n-1$ mots, aucune information d'aucune sorte n'est prise en compte. Cette modélisation ignore donc complètement le fait que ce qu'elle modélise est un langage qui est structuré et qu'il peut exister des dépendances sémantiques entre les phrases [4]. En particulier, la capacité de prédiction d'un mot est fonction de sa place ordinale dans la phrase, indépendamment de son rôle linguistique.

4.1.2 Le modèle *n-gram* de classes

L'usage des modèles de classes tente de répondre à plusieurs des reproches faits au paragraphe 4.1.1, à l'encontre du modèle *n-gram* de mots, en regroupant les mots par classes lexicales [Brown *et al.*, 1992]. Ainsi, les effets de la loi de Zipf sont réduits puisque ce qui a été observé pour une forme lexicale peut être généralisé à l'ensemble des formes de la même classe. Cette classification induit une structure qui permet de prédire un mot, non plus en fonction de l'enchaînement de mots qui lui succédent, mais en fonction de l'observation d'une suite de classes lexicales.

Principe

Si $C(w)$ désigne la classe du mot w, alors un modèle de classes décompose la probabilité d'un mot sachant son historique de la manière suivante pour un modèle *n-gram* :

$$P(w_i|\Phi(w_{i-n+1},...,w_{i-1})) = P(w_i|C(w_i))P(C(w_i)|\Phi(w_{i-n+1},...,w_{i-1})) \tag{4.6}$$

L'hypothèse la plus courante est qu'un mot ne peut appartenir qu'à une seule classe. La fonction d'équivalence Φ est ici directement appliquée au modèle *n-gram* et donc écrite comme étant fonction des $n-1$ mots précédents. Les mots étant regroupés par classe, plusieurs choix sont possibles pour Φ faisant intervenir ou pas cette classification. L'usage le plus fréquent est de choisir

$$P(w_i|\Phi(w_{i-n+1},...,w_{i-1})) = P(w_i|C(w_i))P(C(w_i)|C(w_{i-n+1}),...,C(w_{i-1})), \tag{4.7}$$

où chaque mot de l'historique est remplacé par sa classe. La qualité de cette modélisation est conditionnée par le choix de la classification $C(\cdot)$.

[4] Cette dépendance sémantique est due à la continuité thématique du discours et s'illustre par exemple avec des anaphores.

Classification

La méthode la plus utilisée classifie de manière non-supervisée les mots dans un nombre de classes données, afin de maximiser la vraisemblance du corpus d'entraînement pour un modèle bigramme de classe. Ceci est équivalent à minimiser l'information mutuelle moyenne entre les classes. Ce critère est introduit dans l'article [Brown *et al.*, 1992], mais des algorithmes plus performants sont décrits dans les articles [Kneser and Ney, 1993; Martin *et al.*, 1995]. Une autre méthode relatée par exemple dans [Jelinek, 1990] se sert des catégories morpho-synaxiques (Part of Speech) pour classer les mots. Ainsi le rôle linguistique du mot est pris en compte et non pas uniquement sa place dans la chaîne de mots comme le fait le modèle *n-gram* de mots. L'article [Niesler *et al.*, 1998] compare ces deux approches sur le corpus « 1994 ARPA CSR HUB-1 evaluation ». Si les deux méthodes améliorent les résultats par rapport à un modèle de langage trigramme, la classification non supervisée donne des résultats nettement meilleurs que ce soit en terme de perplexité ou de WER.

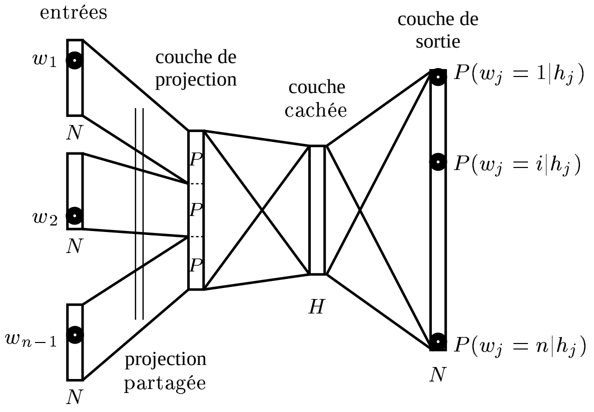
D'autre classifications ont été expérimentées, faisant par exemple appel à des classes sémantiques dédiées à la tâche (comme dans l'article [Ward, 1990] pour un système d'information téléphonique de compagnie aérienne). La conclusion est la même : pour la transcription d'émission radio et télédiffusées, l'état de l'art est un modèle *n-gram* de classes interpolé avec le modèle *n-gram* de mots correspondant, les classes ayant été constituées par la méthode non-supervisée qui minimise l'information mutuelle entre classes.

4.1.3 Modèle connexioniste

La loi de Zipf a comme principale conséquence pour un modèle *n-gram* de mots, le fait que beaucoup de ngrammes ne sont pas ou peu observés lors de l'entraînement, quelle que soit la taille du corpus d'apprentissage. Des techniques ont été développées afin d'attribuer une probabilité non nulle aux ngrammes non-observés lors de l'entraînement (voir le paragraphe 4.4.4). L'objectif est de généraliser les observations faites sur le corpus d'entraînement à ces ngrammes inobservés. Dans dans le cas d'un modèle *n-gram*, l'espace de représentation est l'espace discret des indices des mots. Cet espace est peu propice à cette généralisation, puisqu'il n'existe pas de relation évidente entre les indices[5].

Un modèle de langage connexioniste propose de remédier à ce problème en traitant la tâche d'estimation des ngrammes dans un espace continu. Pour plus de détails sur les modèles de langage connexionistes, il est possible de se reporter aux articles [Bengio and Ducharme, 2001; Schwenk and Gauvain, 2002]. Dans cette thèse, la description d'un modèle connexioniste s'appuie sur le modèle décrit dans l'article [Schwenk and Gauvain, 2002]. L'architecture du modèle de langage connexioniste est représentée à la figure 4.1. Chaque mot est projeté dans un espace vectoriel de dimensions 50, la probabilité d'un mot en contexte étant ensuite estimée à partir de cette représentation. L'utilisation d'un réseau de neurones permet d'apprendre conjointement la projection et l'estimation des densités.

[5]Le fait que deux mots aient par exemple deux indices adjacents, n'apporte aucune information sur la relation qui peut exister entre ces deux mots.



N	:	taille du vocabulaire (65k)
p	:	taille de la projection (50)
H	:	taille de la couche cachée (100-400)
h_j	=	contexte $w_{j-n+1}, ..., w_{j-1}$
$P(w_j = i\|h_j)$		probabilité n-gramme pour le contexte h_j

Tab. 4.1 – Architecture du modèle de langage connexioniste. Les entrées du réseau sont les indices des $n-1$ mots précédents $h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$, et les sorties sont les probabilités *a posteriori* de tous les mots du vocabulaire : $P(w_j = i|h_j), \forall i \in [1, N]$. Puisque le modèle de langage connexioniste possède une complexité plus importante qu'un modèle avec repli standard, le réseau de neurones est souvent utilisé uniquement pour prédire les probabilités des 2000 à 8000 mots les plus fréquents. Les probabilités des autres mots sont calculées par un ML avec repli classique. Le réseau est appris de manière à minimiser la perplexité sur les données d'apprentissage.

Le modèle de langage neuronal permet une meilleure utilisation d'une quantité limitée de données d'apprentissage. Il est principalement utilisé avec succès dans les systèmes de transcription automatique de parole conversationnelle, pour lesquels, la quantité de données d'apprentissage est très limitée[6]. Mais la représentation continue des mots permet également d'envisager des nouvelles techniques d'adaptation non-supervisée qui visent à transformer les mots dans cet espace continu, par exemple pour approcher des mots qui apparaissent souvent dans les mêmes contextes dans une conversation.

4.1.4 Autres modèles

Il existe d'autres types de modèles. Certains sont évoqués à la section 4.5, car ils utilisent comme fondement un modèle *ngram* de mots, comme les modèles *cache* ou *trigger*. D'autres ne sont pas abordés dans ce manuscrit, car ils sont peu utilisés dans le cadre de la reconnaissance de la parole à grand vocabulaire, comme par exemple les modèles basés sur les arbres de décision [Jelinek, 2000].

4.1.5 Faiblesses de la modélisation statistique du langage

Les modèles de langages sont sensibles au changement de style, de thèmes et de genre par rapport au corpus d'entraînement [Lefevre *et al.*, 2001]. Selon R. Rosenfeld [Rosenfeld, 2000], pour

[6]Les transcriptions de vraies conversations sont très rares et très coûteuses.

modéliser les conversations téléphoniques, il est préférable d'utiliser 2 M de mots provenant de transcriptions manuelles de conversations téléphoniques plutôt que 140 M de mots provenant de transcriptions d'émission d'actualité[7]. Pour illustrer ce propos, regardons les dix quadrigrammes les plus fréquents observés sur trois corpus différents : dans le tableau 4.2 les quadrigrammes sont observés sur un corpus d'émission d'actualité de la chaîne CNN des six premiers mois de l'année 2000 et 2003, dans le tableau 4.3 sur un corpus de conversation téléphonique (Switchboard Celular), et dans le tableau 4.4 sur un corpus de textes de journaux (Wall Street Journal). Chacun de ces trois tableaux montre des quadrigrammes spécifiques aux corpus, et la comparaison du tableau 4.2(CNN) avec le tableau 4.3 (conversation téléphonique) appuie les propos de R. Rosenfeld.

Pour le corpus d'émission d'actualité de la chaîne CNN, les dix quadrigrammes les plus fréquents sont conditionnés par l'actualité. Ceci caractérise une des difficultés de la transcription d'émission d'actualité : le contenu linguistique des émissions évolue avec l'actualité. Et ce faisant, l'époque des corpus d'entraînement a un impact sur les performances du modèle [Adda-Decker *et al.*, 1999]. Le tableau 4.2 permet de comparer les quadrigrammes les plus fréquents, observés en 2000 puis en 2003. Certains sont inhérents au type d'émission (« THANK YOU VERY MUCH » ou « WHEN WE COME BACK ») ou au pays (« IN THE UNITED STATES » ou « OF THE UNITED STATES »), et les autres sont liés à l'actualité. Dans le cas des conversations téléphoniques, les quadrigrammes observés dans le tableau 4.3 sont en majorité constitués de mots très fréquents et d'expressions idiomatiques. Alors que pour le corpus Wall Street Journal, la tendance économique du journal est bien représentée dans le tableau 4.4.

Dans l'article [Brill *et al.*, 1998], les auteurs adaptent le jeu de Shannon [Shannon, 1951] afin d'analyser quelles ressources sont sollicitées par un être humain pour corriger des transcriptions automatiques. Trois corpus de parole (conversation téléphonique, transcription d'émissions d'actualité radio et télédiffusées, parole lue) ont été d'abord transcrits automatiquement. Les dix meilleures hypothèses[8] du système de reconnaissance automatique de la parole sont présentées au sujet qui doit, dans un premier temps choisir parmi les dix hypothèses celle qui lui semble la plus juste (correction contrôlée) et dans un second temps, proposer librement une solution. Les améliorations relatives les plus importantes par rapport aux résultats de la transcription automatique sont observées avec la correction contrôlée montrant ainsi l'opportunité d'un post-traitement des 10 meilleurs hypothèses. Le choix libre améliore encore un peu les résultats sans pour autant atteindre ceux de l'oracle[9]. Cependant, les expérimentations ont été faites de manière à ne pas dégrader le WER : en donnant le rang des dix meilleures hypothèses accordé par le système, le sujet choisit, en cas d'hésitation, l'hypothèse du système de reconnaissance.

Dans l'article [Mangu *et al.*, 2003], une expérience similaire est relatée sur un corpus de parole conversationnelle. Les sujets corrigent désormais la sortie du système de reconnaissance automatique de la parole en faisant un choix contrôlé dans un réseau de confusion (également appelé saucisse ou *sausage*), selon deux configurations : lecture seule, et lecture plus écoute du signal.

[7]For example to model casual phone conversations, one is much better off using 2 million words of transcripts from such conversations than using 140 million words of transcripts from TV and radio news broadcast.

[8]Les dix meilleures hypothèses désignent les dix hypothèses considérées comme les plus probables par le système de reconnaissance

[9]le choix parmi les 10 meilleures hypothèses qui minimisent le WER

2000	2003
OF THE UNITED STATES	WEAPONS OF MASS DESTRUCTION
THE FLORIDA SUPREME COURT	THANK YOU VERY MUCH
THANK YOU VERY MUCH	U. N. SECURITY COUNCIL
IN THE UNITED STATES	OF THE UNITED STATES
PRESIDENT OF THE UNITED	FOUR HUNDRED FORTY ONE
U. S. SUPREME COURT	IN THE UNITED STATES
WHEN WE COME BACK	THE UNITED STATES AND
AND I THINK THAT	AT THE WHITE HOUSE
THE U. S. SUPREME	THE U. N. SECURITY
THE SECRETARY OF STATE	ONE THOUSAND FOUR HUNDRED

TAB. 4.2 – Les dix quadrigrammes les plus fréquents observés sur un corpus de transcriptions d'émissions diffusées par CNN en 2000 et en 2003.

YOU_KNOW WHAT I MEAN
OR SOMETHING LIKE THAT
I DO_NOT KNOW I
I DO_NOT KNOW IF
I DO_NOT I DO_NOT
WHAT DO YOU DO
YOU WORK FOR OMNIPOINT
AND STUFF LIKE THAT
YOU_KNOW WHAT I_AM SAYING
DO YOU WORK FOR

TAB. 4.3 – Les dix quadrigrammes les plus fréquents observés sur un corpus de conversations téléphoniques (Switchboard Cellular).

NEW YORK STOCK EXCHANGE
AND CHIEF EXECUTIVE OFFICER
POINT FIVE MILLION DOLLARS
IN NINETEEN NINETY ONE
CHAIRMAN AND CHIEF EXECUTIVE
FIVE CENTS A SHARE
IN NINETEEN NINETY TWO
IN NINETEEN EIGHTY NINE
DOLLARS AND FIFTY CENTS
THE NEW YORK STOCK

TAB. 4.4 – Les dix quadrigrammes les plus fréquents observés sur un corpus de textes du Wall Street Journal.

L'objectif est de déterminer s'il est utile de focaliser les recherches sur des modèles linguistiques plus sophistiqués (au niveau morphologique, syntaxique, sémantique ou pragmatique) ou plutôt sur les modèles acoustiques. Les résultats et les conclusions diffèrent de ceux de l'article [Brill *et al.*, 1998]. Lorsque le sujet ne fait que lire le réseau de confusion, les corrections qu'il effectue dégrade significativement le WER, et seule la lecture et l'écoute permettent de corriger quelques erreurs et donc de réduire le WER. Les auteurs en concluent qu'il est peu probable d'obtenir des gains sur des corpus de parole conversationnelle en améliorant seulement les modèles de langage et que par contre l'amélioration des modèles acoustiques semble plus prometteuse. De plus ils s'interrogent sur les liens entre modèles acoustiques et linguistiques : « Do people do it all together ? ».

4.2 Combinaison de modèles et introduction de contraintes

Cette section décrit deux opérations très fréquentes en modélisation du langage : la combinaison de modèle et l'introduction de contraintes. Lorsque plusieurs sources de connaissances (ou plusieurs corpus) interviennent dans la construction d'un modèle, il peut être utile de construire un modèle pour chacune de ces sources de connaissances, puis de combiner ces modèles en un seul. Cette approche permet de pondérer l'importance de chacune des sources lors de la combinaison des modèles. La méthode la plus utilisée pour cela est l'interpolation linéaire. La seconde opération consiste à construire un modèle unique en imposant des contraintes. Ces contraintes peuvent être des connaissances *a priori* de tout ordre (linguistique, sémantique, thématique, ...). Pour cette opération, l'approche la plus employée est celle du maximum d'entropie.

Mais la définition de ces deux opérations peut paraître arbitraire tant elles sont proches l'une de l'autre en pratique. Par exemple, l'introduction d'une contrainte peut se faire en interpolant linéairement deux modèles, le premier étant le modèle à contraindre, et le second le modèle représentant la contrainte. Inversement, la combinaison de plusieurs modèles peut être reformulée selon un ensemble de contraintes qui sera résolu grâce au principe du maximum d'entropie. Dans ce chapitre, beaucoup d'exemples illustreront ce propos : par exemple les modèles *cache* ou *trigger* sont des contraintes qui s'appliquent généralement à un modèle de langage *n-gram* soit via le maximum d'entropie soit par interpolation linéaire.

Cependant, l'interpolation linéaire a de nombreux avantages. Elle est simple et rapide à mettre en œuvre, et elle permet de manipuler les modèles sans les réestimer. Au contraire, l'application du maximum d'entropie se fait à l'aide d'algorithmes itératifs coûteux et nécessite la réestimation complète du modèle.

4.2.1 Interpolation linéaire

En pratique un modèle de langage est la plupart du temps obtenu à partir d'un mélange de modèles, où chaque constituant a son propre corpus d'entraînement (voir à la section 4.4). L'interpolation linéaire permet de combiner M modèles différents représentés par leur distribution P_i avec $i = 1, \ldots, M$, dont le résultat est le modèle $P_{interpol}$. Elle se définit comme il suit :

$$P_{interpol}(w|h) = \sum_{i=1}^{M} \lambda_i P_i(w|h),$$
$$\text{avec } \sum_{i=1}^{M} \lambda_i = 1. \tag{4.8}$$

Le vecteur $(\lambda_i)_{i=1}^{M}$ désigne les coefficients d'interpolation qui pondèrent l'importance de chaque modèle. Leur estimation se fait généralement en maximisant la vraisemblance d'un corpus de développement au moyen de l'algorithme E.M. [Dempster *et al.*, 1977].

4.2.2 Maximum d'entropie

Alors que l'interpolation linéaire mélange des modèles préalablement construits, l'approche du maximum d'entropie consiste à construire un seul modèle qui inclut toutes les informations données par les différentes sources d'informations. Pour cela, chaque corpus, ou chaque source d'information induit un ensemble de contraintes sur la distribution finale. L'intersection de toutes les contraintes, si elle n'est pas vide, contient un ensemble de distributions qui satisfait à toutes les contraintes. L'approche du maximum d'entropie permet de faire un choix parmi cet ensemble en ne faisant aucune hypothèse sur ce qui est inconnu. L'absence de connaissance se traduit sur un modèle statistique par une distribution uniforme, cette source atteint alors son entropie maximale. La méthode du maximum d'entropie consiste donc à modéliser chaque connaissance par une contrainte, et tout ce qui n'est pas contraint est déterminé de manière à maximiser l'entropie. L'algorithme *GIS* (*Generalized Iterative Scaling*) est le plus fréquemment employé pour résoudre ce problème, et la solution est une famille de modèles souvent appelée modèle exponentiel.

De plus amples informations peuvent-être trouvées dans l'article [Berger *et al.*, 1996] qui décrit l'approche du maximum d'entropie appliquée au traitement du langage naturel en général et à la traduction automatique en particulier. Pour ce qui est de son application à la modélisation du langage pour la reconnaissance automatique de la parole, dans l'article [Pietra *et al.*, 1992], les auteurs introduisent la méthode d'adaptation *MDI* : *minimum discriminant information* qui est équivalente à l'application du principe du maximum d'entropie. Dans l'article [Rosenfeld, 1996] ou l'ouvrage [Jelinek, 2000] plusieurs exemples d'application sont détaillés comme le mélange de modèles ou les modèles *cache* et *trigger*.

4.3 Évaluation des modèles de langages

Évaluer un modèle de langage, c'est évaluer sa contribution dans la tâche qui nous préoccupe : la reconnaissance automatique de la parole. La mesure consacrée est le taux d'erreurs sur les mots, abrégé par WER (voir paragraphe 3.2.2). Cette mesure nécessite cependant de disposer d'un corpus audio de test et de sa transcription manuelle de référence. Le WER est calculé sur les résultats du système de reconnaissance qui fait intervenir la modélisation linguistique mais aussi la modélisation acoustique. Cette mesure requiert un temps de calcul important, puisqu'il faut effectuer complètement le processus de reconnaissance de la parole. D'autre mesures, moins coûteuses, ont été alors proposées. La plus utilisée, et pratiquement la seule, est la perplexité introduite en 1977 dans [Bahl *et al.*, 1977].

4.3.1 Perplexité

L'équation originelle 3.2 sépare la contribution de chacun des éléments du système de reconnaissance. Le modèle de langage participe à la recherche de la meilleure hypothèse en fournissant au décodeur la probabilité *a priori* des hypothèses. Un modèle de langage peut donc être évalué en lui demandant quelle probabilité il assigne aux bonnes hypothèses. Cela revient à constituer un corpus textuel T représentatif de ce que l'on souhaite transcrire, et à calculer avec le modèle M la log-vraisemblance $\mathcal{L}$ moyenne de ce corpus comme suit :

$$\begin{aligned} \mathcal{L}(\boldsymbol{T}|M) &= \frac{1}{n}\sum_{i=1}^{n}\log_2 P_M(w_i|h_i) \\ ppx_M(\boldsymbol{T}) &= 2^{\mathcal{L}(\boldsymbol{T}|M)} \end{aligned} \tag{4.9}$$

La perplexité du texte T selon le modèle M est notée $ppx_M(\boldsymbol{T})$. Elle s'obtient à partir de la log-vraisemblance comme le montre l'équation 4.9. Le corpus de texte $\boldsymbol{T}$ est la suite de mot $\boldsymbol{T} = w_1, w_2, ..., w_n$. Le modèle M assigne la probabilité $P_M(w|h)$ au mot w lorsqu'il est précédé de l'historique h.

Une autre approche similaire qui utilise la théorie de l'information permet de définir la perplexité comme une approximation de l'entropie croisée $\mathcal{C}$ de la distribution inconnue $P(\cdot)$ et sa modélisation M. La perplexité $ppx(P, P_M)$ s'obtient alors en exprimant l'entropie croisée en nombre de bits nécessaire pour la coder.

$$\begin{aligned} \mathcal{C}(P, P_M) &= \textstyle\sum_T P(w_i)\log_2 P_M(w_i|h_i) \\ ppx(P, P_M) &= 2^{\mathcal{C}(P, P_M)} \end{aligned} \tag{4.10}$$

La fonction $\mathcal{C}$ est une approximation car la distribution $P(\cdot)$ est à distinguer de la source ergodique qu'elle est censée modéliser [Manning and Schütze, 1999][Cover and Thomas, 1991]. La perplexité peut s'interpréter comme le facteur de branchement moyen (au sens géométrique) du langage avec le modèle considéré. Le facteur de branchement [Goodman, 1976] se définit habituellement dans un contexte déterministe par le nombre moyen de mots pouvant apparaître à la suite d'une séquence de mots. Cette mesure permet de mesurer les contraintes qu'exerce la grammaire sur le langage. La perplexité en est l'équivalent probabiliste. Si le modèle de langage n'est qu'une distribution équiprobable, tous les mots sont alors également possibles et la perplexité est au maximum : N qui est la taille du vocabulaire.

La perplexité est autant utilisée que criticable. Sa principale qualité est de permettre une évaluation facile et indépendante de toute composante acoustique. Elle est très répandue pour cela. Mais pour servir d'élément de comparaison entre différents modèles, il est préférable que le texte $\boldsymbol{T}$ soit le même. De plus, si les modèles de langage sont construits avec des vocabulaires différents, la perplexité doit être interprétée avec précaution. En effet, lorsque le taux de mots inconnus à l'apprentissage est élevé, les mots inconnus se projettent tous sur l'unique forme <UNK>. Ces mots prennent un poids élevé et la probabilité $p(w_i|h_i)$ est surévaluée par la présence de mots inconnus parmi (w_i, h_i). Mais, *in fine*, là où la perplexité est la plus criticable, c'est qu'elle n'est pas directement liée au WER.

4.3.2 Perplexité *vs* WER

Nombreux sont les articles où les auteurs constatent que « la réduction de la perplexité ne se traduit pas par un gain équivalent sur le WER ». Ainsi dans les articles [Clarkson and Robinson, 1998], [Clarkson and Robinson, 1999], et [Clarkson and Robinson, 2001], différents modèles de langage sont construits, ils conduisent à la même perplexité sur des données de test (1996 Hub4 developpment test) mais bien entendu pas au même WER. Dans [Chen *et al.*, 1998], 35 modèles de langages sont estimés qui se répartissent à peu près équitablement selon deux catégories : la première regroupe des modèles *n-grams* entraînés sur des données du domaine à savoir des transcriptions manuelles d'émissions radio-télédiffusées, la seconde fait varier d'une part les corpus d'apprentissage en introduisant des corpus hors-domaines (Switchboard et American Press), et d'autre part le type de modèle (modèle de classes, de cache, trigger et mélange de modèle). Les conclusions de l'étude de la corrélation entre la perplexité et le WER est similaire à celle de l'étude relatée dans l'article [Iyer *et al.*, 1997] pour la tâche Switchboard (Conversation téléphonique). La perplexité se révèle être un indicateur fiable du WER pour des modèles *n-grams* entraînés sur des données du domaine, ce qui n'est plus le cas lorsque des données hors-domaines sont utilisées.

Ainsi les articles précédemment cités proposent des mesures alternatives. Si les auteurs de l'article [Chen *et al.*, 1998] semblent déçus par leurs propositions, ceux de l'article [Clarkson and Robinson, 2001] émettent l'hypothèse qu'une des faiblesses de la perplexité est qu'elle sollicite uniquement les probabilités des mots présents dans les textes de test. Les auteurs proposent donc de prendre également en compte (grâce à trois mesures) la distribution des probabilités des mots qui sont en compétition avec le mot correct pour un historique donné. La combinaison de ces trois mesures crée un critère, qui est apparemment un indicateur plus fiable du WER. Ce critère semble également plus approprié que la perplexité pour calculer les coefficients d'interpolation d'un mélange de modèles.

Ces expériences incitent donc à manipuler avec précaution la perplexité comme critère de comparaison entre différents modèles. Les nombreux paramètres qui entrent en jeu dans la construction d'un ML ont une incidence sur la validité de la perplexité. Ronald Rosenfeld dans [Rosenfeld, 2000] propose cependant une règle drastique qui s'applique à la transcription automatique de documents audiovisuels : une réduction de la perplexité par rapport à un « bon modèle de référence » n'est pas significative lorsqu'elle est inférieure ou égale à 5%, elle est digne d'intérêt entre 10 et 20% et laisse éventuellement présager une réduction du WER, et au-delà de 30%, c'est une réduction aussi significative que rare.

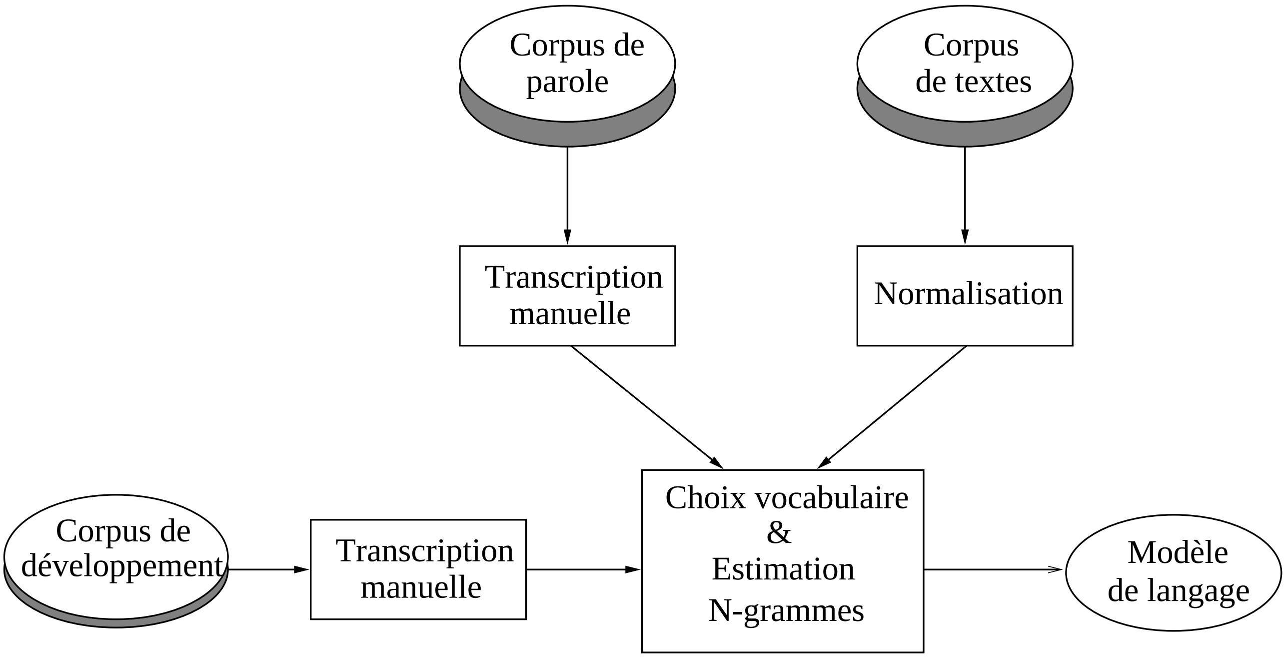


FIG. 4.4 – Processus de construction d'un modèle de langage

4.4 Construction du modèle *n-gram* de référence

La construction d'un modèle *n-gram* de mots ne se résume pas à estimer des probabilités à partir d'un corpus d'entraînement en utilisant un toolkit comme celui de S.R.I [Stolcke, 2002] ou celui de C.M.U-Cambridge [Clarkson and Rosenfeld, 1997]. Tout le travail préalable, tel que le choix des corpus, leurs utilisations et leurs normalisations, conditionne les performances du modèle. La présente section explicite chacune des étapes de la construction d'un modèle de langage telle qu'elle est représentée par le schéma de la figure 4.4, et en prenant pour exemple celui qui est par la suite appelé le modèle de langage de référence. Ce modèle a été construit en 2000 selon les méthodes décrites dans les articles [Adda-Decker *et al.*, 1999; Adda *et al.*, 1997a]. La démarche est sensiblement la même que celle suivie pour l'anglais [Gauvain *et al.*, 2002].

4.4.1 Description des corpus

Les textes disponibles au moment de la construction du modèle de référence se répartissent selon 5 sources. La première contient les transcriptions fines de journaux radio-télévisés qui ont été utilisés lors de l'apprentissage des modèles acoustiques. Ces émissions proviennent de 6 diffuseurs et se répartissent sur 6 années de 1994 à 1999. La deuxième source contient des transcriptions d'extraits d'émission datant des années 1997, 1998 et 2000, réalisées par une agence de service de presse. Les trois dernières sources proviennent de la presse écrite. Ce sont des dépêches d'agences de presse datant de 1994 à 1996, des articles du journal *Le Monde* de 1987 à 1998, et des articles du journal *Le Monde Diplomatique* de 1990 à 1996.

Source	*Période*	*Nombre de mots*
Transcriptions	1994-1999	1 603 800
Services de presse	1997,1998,2000	72 653 790
Agence de presse	1994-1996	66 815 080
Le Monde	1987-1998	257 406 060
Le Monde Diplo.	1990-1996	6 714 390
Total	1987-2000	405 193 120

TAB. 4.5 – Chaque source est caractérisée par la période couverte et le nombre total de mots.

Le tableau 4.5 résume la quantité de données que contient chaque source ainsi que les périodes couvertes. Le nombre de mots est calculé après l'étape de normalisation qui est décrite au paragraphe 4.4.2.

4.4.2 Définition d'un mot et normalisation des textes

La normalisation est une étape majeure dans la préparation du corpus d'entraînement qui consiste à segmenter les textes en phrases et en mots [Manning and Schütze, 1999]. C'est à partir de la forme normalisée des textes d'entraînement, que le vocabulaire et le modèle de langage du système de reconnaissance sont construits. Les performances du système sont donc conditionnées par cette étape. Les corpus de textes décrits au paragraphe précédent 4.4.1 ont été traités par Gilles Adda. Les articles [Adda and Adda-Decker, 1997] et [Adda *et al.*, 1997b] décrivent en détails la méthode appliquée et les choix qui ont été effectués ainsi que les impacts de la normalisation sur les performances. Tout texte utilisé pour la construction d'un modèle passe par ces transformations à l'exception des textes extraits de sites Web (Pour ces derniers, se référer à la section 6.1.1 qui leur est consacrée).

La segmentation en phrases et en mots doit satisfaire deux critères contradictoires : obtenir un vocabulaire ayant la meilleure couverture possible, et définir les mots de manière non-ambiguë. Le premier critère a pour objectif de diminuer les erreurs de reconnaissance dues aux mots hors vocabulaire et incite à réduire le nombre de formes différentes en uniformisant la capitalisation et en séparant toutes les ponctuations des mots (éclatement des mots composés, des sigles et des acronymes). Il en résulte une grande ambiguïté syntaxique induite par la création d'homographes comme par exemple « Roman » (Polanski) et « roman », ou « sec. » (abréviation de seconde) et « sec ». *A contrario*, le second critère incite à ne rien transformer, multipliant les formes distinctes, mais parfois de façon peu pertinente comme c'est le cas avec « c'est », « c' est » et « C'est ». La normalisation implique donc de définir précisément ce qu'est un mot et une phrase.

La capitalisation

L'usage en anglais est de ne pas distinguer les majuscules des minuscules. Cette pratique n'a pas d'autre justification que l'habitude prise par la communauté après les premières campagnes d'évaluation américaine sur la reconnaissance automatique de la parole. En français, la lettre majuscule est principalement utilisée en début de mot pour marquer un début de phrase, un nom propre (éventuellement une antonomase), ou les deux. Elle compose également les sigles et les acronymes. La distinction est donc maintenue sauf dans le cas où elle marque un début de phrase.

Définition d'un mot

Dans le contexte de cette thèse, un mot se définit par son orthographe[10]. Ceci implique que deux inflexions ou dérivations de la même racine verbale sont deux mots distincts (par exemple pour le verbe être : « est » et « es », et pour beaucoup de noms communs la différence de nombre crée deux mots comme « fromage » et « fromages »). Paradoxalement, deux homographes ayant des sens différents ou des fonctions différentes sont représentés par une seule forme (par exemple « est » suivant qu'il soit un nom ou la forme conjugée du verbe être, et « livre »).

Différents termes sont couramment utilisés comme mot, forme, forme lexicale. Tous ont le même sens, et se réfèrent au terme anglais *token* : une suite de caractères alpha-numériques bornée par des caractères de séparation. L'ensemble des caractères de séparation contient l'espace (incluant l'espace multiple et la tabulation), le retour à la ligne ainsi que certaines ponctuations. Il paraît évident de choisir l'espace comme caractère de séparation, même si dans certains cas, assez rares il est vrai, certains mots peuvent s'écrire avec un espace : « New York » et « San Francisco ». Le point est un caractère dont les multiples fonctions posent problème. En effet, s'il marque la fin de phrase, il sert aussi à séparer les lettres des acronymes et des sigles, ou à marquer une abréviation. Certaines de ces fonctions peuvent se superposer et donner ainsi une haplologie comme par exemple dans le cas d'une abréviation en fin de phrase où le point n'est pas répété. Ce problème peut être traité en établissant un ensemble de règles.

Les mots composés

Nous nommerons mot composé, un mot contenant un tiret. La question est de décider s'il est nécessaire de séparer un mot composé ou pas. Séparer tous les mots composés améliore la couverture lexicale car les mots qui constituent un mot composé sont fréquents pour la plupart. Mais séparer les compositions usuelles, telles que « après-demain » ou « Radio-France », altère la capacité discriminante du modèle de langage. Paradoxalement, conserver toutes les formes composées, dégrade considérablement la couverture lexicale en augmentant le nombre de formes sans pour autant apporter une information pertinente (par exemple « Paris-Bordeaux »). Le compromis s'effectue en conservant les compositions usuelles et donc fréquentes, et en séparant les compositions plus rares.

[10]La définition d'un mot est lié au domaine. Par rapport aux nombreuses définitions données par le dictionnaire « le trésor de la langue française informatisé », la définition adoptée dans le cadre de cette thèse est la définition informatique : suite finie et totalement ordonnée de caractères. Ce dictionnaire est consultable à l'adresse http ://atilf.inalf.fr/.

Sigles et acronymes

Un sigle est une suite de lettres initiales constituant l'abréviation de plusieurs termes formant une unité de dénomination fréquemment employée[11]. Un sigle se prononce en l'épelant (par exemple SNCF et RFI). Alors qu'un acronyme se distingue d'un sigle par le fait qu'il se prononce comme un mot (par exemple URSSAF) ; à ce titre, il est considéré comme un mot et a son entrée dans le lexique de prononciation. Dans les textes, il n'y pas en pratique de règles typographiques qui permettent de différencier un sigle d'un acronyme. Cette classification est faite à partir d'une liste d'acronymes et à partir de règles permettant de déterminer si un sigle est prononçable. Pour les sigles se pose une alternative similaire à celle de la séparation des mots composés. La solution est également similaire : un ensemble de sigles fréquents et répertoriés n'est pas modifié, tandis que les autres sont éclatés.

Unités fréquentes et chiffres

Le cas des unités est particuliérement épineux. Pour chaque unité, l'objectif est de l'identifier et de la réécrire dans une forme unique. La difficulté est de répertorier les diverses formes que peut prendre une unité comme par exemple la seconde que l'on rencontre sous différentes formes comme : « sec », « sec. », « s », « " ». Quant aux chiffres, ils sont pré-traités (passage de « 10 000 » à « 10000 ») puis traduits en mots.

Définition d'une phrase

Les étapes précédentes permettent une segmentation en mots. La dernière étape consiste à définir ce qu'est une phrase. Pour un modèle *n-gram* classique, cette unité est importante puisqu'elle borne l'historique. Une phrase peut se définir récursivement par rapport au mot : à l'exception des incises, une phrase est une suite de mots bornée par des caractères de séparation. L'ensemble des caractères de séparation est { **.** , **:** **¨** , **!** , **?** et **...**.

Corpus d'entraînement

Une fois la segmentation en phrases effectuée, les ponctuations sont retirées et la majuscule en tête de phrase est traitée. Ainsi une phrase devient une suite uniquement composée de mots qui sont séparés par des espaces. Un corpus d'entrainement contient un ensemble de phrases encadrées par une balise de début et de fin (respectivement <s> </s>).

4.4.3 Estimation des probabilités

Un modèle de langage *n-gram* attribue la probabilité conditionnelle $P(w|h)$ au mot w et sachant son historique h. h désigne les $n-1$ mots qui précèdent w. Chacun de ces mots appartient

[11] Définition empruntée au « trésor de la langue française informatisé » (http ://atilf.inalf.fr/).

Kandinski et	Brahem	sympathisaient	à Paris ou à	Knock-le-Zoute
Kandinski et	<UNK>	sympathisaient	à Paris ou à	<UNK>

TAB. 4.6 – Exemple de projection d'une phrase sur le vocabulaire. Les mots *Kandinski*, *sympathisaient*, *et*, *à*, *Paris*, et *ou* sont supposés faire partie du vocabulaire, alors que les autres non.

au même ensemble fini qui est le vocabulaire V. La normalisation des probabilités exige pour un historique donné h :

$$\sum_{w \in V} p(w|h) = 1 \tag{4.11}$$

Les probabilités sont estimées à partir d'observations faites sur un corpus d'entraînement. Le vocabulaire étant nécessairement fini, certains mots du corpus d'entraînement peuvent ne pas faire partie du vocabulaire. Aussi le vocabulaire contient une forme particulière (<UNK>), qui peut-être vue comme une classe regroupant tous les mots inconnus (c'est-à-dire qui ne sont pas dans le vocabulaire). Les observations sont obtenues après projection de chaque mot du corpus d'entraînement sur le vocabulaire comme le montre l'exemple 4.6.

Les probabilités observées sont déterminées afin de maximiser la vraisemblance du corpus d'entraînement. La log-vraisemblance du corpus d'entraînement T s'écrit :

$$\mathcal{L}(T) = \sum_h \sum_w N_T(h, w) \log \tilde{P}(w|h) \tag{4.12}$$

où $N_T(h, w)$ désigne le nombre d'occurrences du mot w précédé de son historique h dans le corpus T et $\tilde{P}(w|h)$ la probabilité observée car estimée à partir des observations. La somme s'effectue sur l'ensemble des h et w engendrés par le vocabulaire. Il peut être démontré que l'estimateur qui maximise la vraisemblance est simplement la fréquence relative (par exemple dans [Beaujard, 2000]) :

$$\tilde{P}(w|h) = \frac{N_T(h, w)}{N_T(h)} \tag{4.13}$$

où $N_T(h)$ désigne le nombre d'occurrences de l'historique h dans le corpus T. Le problème de cette estimation est que lorsqu'un événement est inobservé, sa fréquence relative et donc sa probabilité est nulle. Un mot du vocabulaire qui n'apparaît pas dans les textes d'entraînement s'appelle un *zéroton*.

4.4.4 Lissage ou *smoothing*

L'estimation des probabilités à partir des observations n'est pas suffisante, car il n'existe pas de corpus assez grand pour observer tous les ngrammes valides qui peuvent être générés par le vocabulaire, en effet ce n'est pas parce qu'un ngramme est inobservé, qu'il est impossible et que donc sa probabilité doit être nulle. De plus, il est très dangereux, lors du décodage, d'avoir une probabilité nulle pour un événement. Des techniques de lissage sont alors employées pour combler les lacunes des observations et permettre au modèle de langage d'attribuer une probabilité à tout ngramme.

La plupart des méthodes peuvent se décomposer en deux étapes, la première prélève une masse de probabilité sur les ngrammes observés, et cette masse est redistribuée dans la seconde étape sur les événements inobservés. Pour le prélèvement, beaucoup de techniques ont été expérimentées (voir par exemle [Chen and Goodman, 1998]), la redistribution se fait généralement soit par interpolation linéaire, soit par repli à l'ordre inférieur (plus connue sous l'appellation *back-off*).

Redistribution par interpolation linéaire

Le méthode introduite dans l'article [Jelinek and Mercer, 1980] est souvent nommée le lissage Jelinek-Mercer. Elle modifie la probabilité observée $\tilde{P}$ en l'interpolant avec le modèle d'ordre inférieur. Une notation plus compacte est adoptée pour les historiques ou les séquences de mots extraites de $\boldsymbol{W}$: $w_{i-n+1}^{i-1} = w_{i-n+1}w_{i-n+2}\dots w_{i-1}$. Ainsi, la probabilité « lissée » est désignée par $P_{smooth}(w_i|w_{i-n+1}^{i-1})$ et s'écrit comme suit :

$$P_{smooth}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}}\tilde{P}(w_i|w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}})P_{smooth}(w_i|w_{i-n+2}^{i-1}) \tag{4.14}$$

Les coefficients d'interpolation peuvent-être calculés afin de maximiser la vraisemblance, soit d'une partie du corpus d'entraînement qui a été retirée pour l'estimation de $\tilde{P}$ (*held-out* interpolation), soit sur un corpus de développement. Le prélèvement de Kneser-Ney qui est proposé dans l'article [H. Ney, 1994], applique le prélèvement absolu aux ngrammes observés avant l'interpolation. Enfin, la technique *Kneser-Ney modified*, introduite dans [Chen and Goodman, 1996] est considérée comme étant l'état de l'art. Elle se distingue de celle de Kneser-Ney en traitant différemment les ngrammes qui apparaissent 1, 2 et 3 fois ou plus.

Back-off

Historiquement, le repli se distingue de l'interpolation linéaire par le fait que pour estimer la probabilité d'un ngramme observé, les distributions d'ordres inférieurs n'interviennent pas. Dans l'équation 4.15, la probabilité « lissée » est fonction de la probabilité estimée $P(w_i|w_{i-n+1}^{i-1})$, si le nombre d'apparitions de l'historique $N(w_{i-n+1}^{i-1}) > 0$ dans le corpus d'apprentissage, sinon le modèle se replie sur la probabilité de l'ordre inférieur, lissée elle aussi, et pondérée par $\alpha(w_{i-n+1}^{i-1})$. La probabilité estimée P est obtenue à partir des probabilités observées $\tilde{P}$ par une technique de prélèvement.

$$P_{smooth}(w_i|w_{i-n+1}^{i-1}) = \begin{cases} P(w_i|w_{i-n+1}^{i-1}) & \text{si } N(w_{i-n+1}^{i-1}) > 0 \\ \alpha(w_{i-n+1}^{i-1})P_{smooth}(w_i|w_{i-n+2}^{i-1}) & \text{si } N(w_{i-n+1}^{i-1}) = 0 \end{cases} \tag{4.15}$$

Le terme $\alpha(w_{i-n+1}^{i-1})$ est le coefficient de back-off qui remplit deux rôles : d'abord il pénalise le recours au repli, et il assure la normalisation des distributions (équation 4.11). Cette méthode a été introduite dans l'article [Katz., 1987], en utilisant le prélèvement de Good-Turing [Good, 1953]. Elle a également été adoptée avec le prélèvement de Witten-Bell [Witten and Bell, 1991].

La distinction entre le repli et l'interpolation linéaire est pratique et historique. D'un point de vue théorique, il est toujours possible par un changement de variable et le choix de l'estimateur P de se ramener au formalisme du repli [Kneser and Ney, 1995], décrit par l'équation 4.15. Deux références au même titre et des mêmes auteurs sont très utiles pour aborder le lissage [Chen and Goodman, 1996] et [Chen and Goodman, 1998]. La seconde est un rapport technique qui étend considérablement le propos de la première. Ces deux articles mènent, entre autres, une évaluation des techniques de lissage. Les auteurs concluent sur le fait que la technique de lissage la plus performante est *Kneser-Ney modified*, introduite par eux-mêmes.

4.4.5 Utilisation des corpus

La critique émise au paragraphe 4.1.5 sur la sensibilité des modèles de langages au changement de style, de thèmes et de genre par rapport au corpus d'entraînement, s'applique au domaine de la transcription automatique de documents d'actualité radio-télédiffusés. Le problème est que peu de données du domaine sont disponibles, comme le montre le tableau 4.5 (voir paragraphe 4.4.1). Des journaux comme *Le Monde* traitent en partie les mêmes sujets et les mêmes thèmes liés à l'actualité, mais avec un langage différent puisqu'écrit. La différence de taille entre les différentes sources, empêche de simplement réunir les différentes sources en un seul corpus d'entraînement. Si tel était le cas, les transcriptions manuelles serait totalement noyées dans la masse que représente *Le Monde*. Ainsi il est préférable de constituer plusieurs corpus d'entraînement, qui respectent les spécificités et les proportions de leurs constituants. Un modèle de langage est estimé à partir de chaque corpus ; les modèles sont finalement linéairement interpolés pour donner le modèle de référence.

Des expériences similaires à celles décrites pour l'anglais dans [Gauvain *et al.*, 2002] ont conduit à la constitution de trois corpus par le regroupement suivant : tous les textes de la presse écrite (les journaux et les dépêches d'agences de presse), les services de presse, et enfin les transcriptions manuelles. Trois modèles ont été estimés sur chacun des corpus et ils ont été interpolés linéairement pour donner le modèle de référence auquel on se réfère dans le cadre de cette thèse.

4.4.6 Description du modèle de langage de référence

Le modèle de langage de référence désigne un modèle ngramme issu de l'interpolation linéaire de trois modèles ngrammes avec back-off. Ces trois modèles ont été estimés sur respectivement un corpus de textes de la presse écrite (les journaux *Le Monde*, *Le Monde diplomatique*, les dépêches AFP), un corpus de transcription d'extraits d'émissions télé et radiodiffusées, et un corpus de transcription d'émissions. Les trois utilisent le prélèvement de Witten-Bell. Le modèle quadrigramme résultant contient 15 millions de bigrammes, 13 millions de trigrammes et 10 millions de quadrigrammes[12]. Par la suite, tous les modèles construits seront des modèles ngrammes avec back-off.

[12]Les seuils appliqués sur les fréquences d'apparition des ngrammes sont différents suivant l'ordre.

4.5 Adaptation des modèles de langage

Tous les modèles de langage décrits précédemment se construisent à partir d'observations faites sur des corpus d'entraînement. Pour une application donnée sur un corpus de test (parole lue, document audiovisuel, ou parole conversationnelle), les performances d'un modèle dépendent donc de l'adéquation entre le corpus d'entraînement et le corpus de test. Pour que cette adéquation soit vérifiée, il est d'abord nécessaire de caractériser la tâche en terme de source de langage (ou type de langage), pour pouvoir ensuite constituer un corpus d'entraînement qui permette de la modéliser. Ce travail correspond à la construction d'un modèle de langage telle qu'elle est décrite à la section 4.4.

Le besoin d'adapter un modèle de langage intervient lorsque le modèle se révèle être en inadéquation avec la tâche, et que les données permettant de construire un modèle pour cette tâche ne sont pas disponibles en quantité suffisante. Cette inadéquation peut être due au fait que les données d'entraînement contiennent peu ou pas de données permettant de modéliser la source de langage d'où proviennent les données de test. Dans l'article [Rosenfeld, 1996], l'adaptation est divisée selon trois paradigmes. Le premier envisage le cas où les corpus d'apprentissage et de test proviennent de la même source, mais cette source est subdivisée en « sous-langages » ou en thèmes. L'adaptation consiste alors à spécialiser le modèle de langage à partir des mêmes données d'apprentissage. Les modèles adaptatifs sont généralement ceux utilisés dans ce cas et sont décrits au paragraphe 4.5.1. Les deux autres paradigmes correspondent au cas où le corpus de test provient d'une source non représentée par les données d'entraînement. Les symptômes en sont une augmentation des observations des mots et des ngrammes nouveaux sur le corpus de test. L'adaptation consiste alors en une extension du modèle à cette nouvelle source. Ce cas se divise lui-même en deux paradigmes qui sont décrits aux paragraphes 4.5.2 et 4.5.3, suivant la disponibilité ou pas d'une connaissance *a priori* sur la nouvelle source de langage.

Dans la plupart des cas, l'adaptation passe par une modification des distributions du modèle de référence en fonction d'une distribution cible ou de connaissances *a priori*. Deux techniques introduites à la section 4.2, sont principalement rencontrées pour faire ces modifications. Il s'agit de l'interpolation linéaire, et de l'application du principe du maximum d'entropie (également appelée adaptation MDI). Quasiment toutes les méthodes décrites par la suite peuvent changer de catégorie au gré des articles, car chaque technique peut être utilisée soit pour adapter un modèle de langage soit pour générer un modèle adaptatif, comme les modèles *cache* ou l'adaptation *MDI*.

4.5.1 Les Modèles adaptatifs

Les modèles adaptatifs (ou *adaptive models*) se définissent ainsi car ils se spécialisent dynamiquement au cours du processus de reconnaissance automatique de la parole en fonction de ce que le système a déjà reconnu. Pour la transcription automatique de documents audiovisuels, il est possible d'utiliser une transcription automatique partielle ou complète du document effectuée par le système de reconnaissance. Dans le cas d'un processus de reconnaissance en plusieurs passes (voir le chapitre 3), il est également possible d'utiliser les résultats d'une passe antérieure.

Modèle cache

Le modèle *cache* a été introduit dans l'article [Kuhn and Mori, 1990]. Il part du principe que si un mot vient d'être reconnu, alors sa probabilité d'apparition s'en voit accrue. Ainsi les mots les plus récemment reconnus sont stockés dans un *cache*. Pour chaque mot, les probabilités P_{ref} attribuées par le modèle de langage de référence, sont modifiées en fonction des probabilités unigramme P_{cache} observées sur le *cache*. Dans le cas d'une adaptation par interpolation linéaire, cela s'écrit pour le i-ème mot d'une hypothèse :

$$P(w_i|h_i) = (1 - \lambda)P_{ref}(w_i|h_i) + \lambda P_{cache}(w_i) \tag{4.16}$$

où le coefficient d'interpolation λ peut être calculé afin de maximiser la vraisemblance d'un corpus de développement. Une variante consiste à faire décroître la probabilité P_{cache} de manière exponentielle en fonction de la distance entre la position actuelle et les apparitions précédentes du mot [Clarkson and Robinson, 1997]. L'introduction du *cache* en tant que contrainte par le principe de maximum d'entropie est décrit dans [Jelinek, 2000].

Modèle trigger

L'adaptation de type *cache* modifie uniquement les probabilités des mots observés dans le *cache*. Or une séquence ou un mot qui vient d'être reconnu, peut modifier la probabilité d'apparition d'autres séquences ou d'autres mots qui lui sont généralement associés [Rosenfeld and Huang, 1992]. Ceci revient à modéliser une dépendance entre les mots sur une plus longue distance que les $n-1$ mots précédents, en faisant l'hypothèse que si une séquence de mot A est fortement corrélée à la séquence de mot B, alors l'apparition de la séquence A modifie la probabilité $P(B|A)$ d'apparition de la séquence B sachant A.

La construction d'un modèle *trigger* commence par la sélection des paires de déclenchement (*trigger pairs*). Le critère utilisé est un seuil soit sur les fréquences de co-occurrences soit sur l'information mutuelle entre les deux séquences. Ces deux grandeurs sont estimées à partir d'observations sur un corpus d'entraînement. La plupart du temps, les séquences se limitent à un mot. L'intégration de ces probabilités se fait comme à l'accoutumée, soit par interpolation linéaire avec le modèle de langage de référence, soit en utilisant le principe de maximum d'entropie [Rosenfeld, 1996]. Dans l'article [Beeferman *et al.*, 1997], les auteurs introduisent, comme dans le cas du modèle *cache*, la notion de distance du type exponentielle décroissante entre les séquences.

Mélange de modèles thématiques

Dans cette approche, le corpus d'apprentissage est divisé automatiquement en un nombre donné de composantes. Chacune d'entre elles modélise un sous-langage ou un thème. Un modèle est estimé sur chaque composante, et ces modèles sont interpolés linéairement pour donner le modèle final comme pour le modèle de référence (voir paragraphe 4.4.5). La différence vient du fait que les coefficients d'interpolations sont estimés dans le but de maximiser la vraisemblance d'une hypothèse du système de reconnaissance. Différentes méthodes ont été expérimentées pour la division du corpus d'entraînement comme par exemple dans les articles [Clarkson and Robinson, 1997] ou [Seymore and Rosenfeld, 1997]. Cependant, les conclusions à propos de ces méthodes est qu'il est possible de diminuer la perplexité mais peu le WER.

D'une manière générale, les gains observés par le passé grâce aux modèles adaptatifs se sont réduits au fur et à mesure que les modèles *n-gram* ont été améliorés. En particulier, la spécialisation que permettent les modèles adaptatifs a été compensée par des corpus d'entraînement toujours plus importants dont les modèles *n-gram* ont su tirer profit.

4.5.2 Adaptation non supervisée des modèles de langage

Lorsqu'aucune information sur le corpus de test n'est disponible, il est possible d'acquérir des connaissances *a priori* de son contenu linguistique grâce à sa transcription automatique. Malheureusement, la transcription automatique contient des erreurs, et cela représente peu de données. Les auteurs de l'article [Chen *et al.*, 2001] proposent d'utiliser des techniques de recherche d'informations, qui tentent de généraliser les connaissances contenues dans une transcription automatique préalable tout en limitant l'impact des erreurs de reconnaissance. Les expériences qui y sont relatées portent sur la tâche de transcription automatique d'émissions d'actualité radio et télédiffusées.

Dans cet article, les textes d'entraînement ont été regroupés thématiquement soit automatiquement selon 198 thèmes soit selon 8 thèmes définis manuellement. La transcription automatique est d'abord segmentée en « histoires » (traduction de *story* qui peut se définir comme un segment de texte se caractérisant par l'unicité du thème), puis pour chaque histoire un ensemble de mots-clés est extrait, et à l'aide de ces mots-clés des documents pertinents sont sélectionnés pour constituer un corpus d'adaptation. Ce corpus sert soit à calculer les coefficients d'interpolation d'un modèle adaptatif du type mélange de modèles thématiques comme cela a été décrit au paragraphe précédent 4.5.1, soit à adapter le modèle de référence selon la méthode du maximum *a posteriori* ou MAP [Federico, 1996; Duda and Hart, 1974].

Toujours, dans l'article [Chen *et al.*, 2001], lorsque que les coefficients d'interpolation d'un mélange de modèles thématiques sont estimés directement sur la transcription, les performances de reconnaissance sont dégradées alors que la perplexité est réduite de 10%. Par contre, lorsque les coefficients sont calculés sur les transcriptions de références ou sur le corpus d'adaptation, des gains sont obtenus à la fois en terme de perplexité mais aussi en terme de CER (Character Error Rate qui est l'équivalent du WER pour le Mandarin). Ces résultats tendent à valider l'hypothèse de départ qui incriminait les erreurs de la transcription automatique. Les meilleurs résultats sont obtenus avec l'adaptation MAP : une réduction du CER de 18,5% à 17,7% (une réduction relative de 4,3%). Cependant cette méthode n'utilise pas d'information nouvelle et en particulier garde le vocabulaire constant.

4.5.3 Adaptation utilisant des connaissances *a priori*

Le cas de figure traité au paragraphe 4.5.2 est celui où aucune connaissance *a priori* sur le document à transcrire n'est disponible. L'usage d'une transcription automatique initiale est alors la seule information qui permette de caractériser la source dont sont issues les données de tests par rapport aux données d'entraînement. Mais il existe des cas plus favorables, où l'on dispose de quelques mots ou de quelques phrases concernant le corpus de test. La quantité de ces données d'adaptation est très variable, et c'est un des critères les plus importants qui conditionne le choix des techniques d'adaptation à employer.

Dans le cas des émissions d'actualité, l'acquisition de connaissances *a priori* sur les données de test peut se faire en faisant l'hypothèse que des textes contemporains à l'émission, abordent les mêmes sujets d'actualité et donc les mêmes thèmes. Pour modéliser le contenu lexical et linguistique de l'actualité, un corpus constitué de journaux et de dépêches d'agences de presses couvrant la même période peut-être utilisé. Plusieurs articles ont montré l'impact de l'époque des corpus d'entraînement sur le vocabulaire et les modèles de langages comme [Adda-Decker *et al.*, 1999], [Barras *et al.*, 2002], et [Rosenfeld, 1995]. Une fois la source nouvelle modélisée, le modèle de référence peut être modifié de deux manières : soit seul le modèle de langage est adapté, soit le vocabulaire est d'abord adapté puis le modèle de langage.

Adaptation MDI

Lorsque peu de connaissances *a priori* sont disponibles (12 000 mots par exemple pour [Kneser *et al.*, 1997] et entre 1 000 et 3 000 pour [Federico, 1999]), seul une distribution unigramme $\hat{P}_A(w)$ peut être estimée. Le modèle de langage adapté noté $P_A(w,h)$ est alors obtenu en lui imposant la distribution unigramme $\hat{P}_A(w)$, tout en minimisant l'entropie relative entre $P_A(w,h)$ et $P_B(w,h)$. Ainsi le modèle $P_A(w,h)$ est contraint à la distribution unigramme obtenue à partir des connaissances *a priori* spécifique au document à transcrire, tout en le maintenant proche en terme d'entropie relative, du modèle de langage de référence $P_B(w,h)$ qui lui a été estimé sur des corpus conséquents. Ce processus est décrit par le couple d'équations suivantes :

$$\begin{aligned} P_A(.) &= \underset{P(.)}{\operatorname{argmin}} \sum_{h,w \in V^n} P(w,h) \log\left(\frac{P(w,h)}{P_B(w,h)}\right), \\ \text{avec } \hat{P}_A(w) &= \sum_h P_A(h) P_A(w|h) \ \forall w \in V \end{aligned} \tag{4.17}$$

L'entropie relative se nomme aussi divergence de Kullback-Leibler ou *discriminant information* [Cover and Thomas, 1991]. Le système d'équations 4.17 est typique de l'approche maximum d'entropie dont la solution peut être trouvée via l'algorithme *GIS* (*Generalized Iterative Scaling*). Cet algorithme est amplement décrit dans ce cas précis dans les articles [Federico, 1999], [Kneser *et al.*, 1997]. L'adaptation *MDI* a été introduite dans le cadre d'un modèle adaptatif [Pietra *et al.*, 1992]. Cette méthode est équivalente à l'application du maximum d'entropie.

Dans l'article [Federico, 1999], les expériences ont été menés sur un corpus de journaux télévisés italien de 40 minutes réparties en 10 parties sur deux mois (octobre et décembre 1998). Des textes collectés sur internet et contemporains des dix parties constituent dix corpus d'adaptation d'environ 2 000 mots. Pour chacune des dix parties, la distribution unigramme $\hat{P}_A(w)$ est estimée à partir du corpus d'adaptation et permet de déterminer le modèle adapté $P_A(w|h)$. Le vocabulaire de référence contient 60k mots et conduit à un taux de mots hors vocabulaire de 2% sur les données de test. L'adaptation à vocabulaire constant permet une réduction relative du WER de 5% (de 28,6% à 27,1%), alors qu'il est de 6% lorsque le vocabulaire est augmenté.

Extension du vocabulaire et interpolation linéaire

Les expériences décrites dans l'article [Federico and Bertoldi, 2001] suivent celles de l'article [Federico, 1999] qui vient d'être évoqué. Le corpus de test est toujours constitué de journaux télévisés italiens, mais il est plus récent et d'une durée totale de 6h (l'évaluation porte uniquement sur les segments de parole bande large). Les auteurs proposent un modèle de langage à vocabulaire ouvert qui inclut une classe spéciale MHV [13] ainsi que les contraintes suivantes énoncées pour un modèle trigramme :

$$\begin{aligned} P(x|w_{i-2}w_{i-1}) &= P(x) \\ P(w_i|w_{i-2}x) &= P(w_i) \\ P(w_i|xw_{i-1}) &= P(w_i|w_{i-1}) \end{aligned} \tag{4.18}$$

où w_i désigne un mot appartenant au vocabulaire de 62k mots du modèle de langage et x un mot nouveau. Ainsi l'introduction de mots dans le modèle de langage peut se faire en remplissant la classe MHV. La probabilité unigramme d'un mot nouveau peut soit être fixée en fonction de la taille d'un vocabulaire universel (10^6) soit estimée à partir d'un corpus contemporain composé de textes de journaux et de dépêches d'agences de presse. Trois méthodes de sélection des mots nouveaux sont expérimentées dans cet article, qui utilisent un corpus de textes de journaux contemporain : la première choisit les nouveaux mots les plus récents, la deuxième choisit les nouveaux mots les plus fréquents et la troisième qui combine les deux premières est de loin la meilleure. Cette méthode permet une réduction relative moyenne de 58% du taux de mots hors vocabulaire avec l'ajout de 60k mots (de 1,6% à 0,7%), de 3,3% de la perplexité (301 à 291) et 2,1% sur le WER (de 25,4% à 24,5%). Les mots nouveaux ont été phonétisés automatiquement. Une adaptation supplémentaire des ML utilise l'interpolation linéaire du modèle estimé sur le corpus parallèle avec le modèle utilisé précédemment et conduit à un gain supplémentaire en perplexité de 10%, mais que de 1,3% relatif en WER.

Modèle de langage tournant

Dans l'article [Auzanne *et al.*, 2000], les auteurs mettent en place un modèle de langage « tournant » qui utilise un corpus parallèle pour mettre à jour toutes les semaines le vocabulaire, le dictionnaire de prononciation puis le ML. Le corpus parallèle est nommé ainsi car il contient des textes de journaux d'actualité contemporains des données à transcrire (couvrant les 28 derniers jours). Le vocabulaire de référence contient 50 000 mots qui sont maintenus. Il reste de la place pour environ 15 000 nouveaux mots afin de respecter la contrainte sur la taille du vocabulaire. Les nouveaux mots sont sélectionnés à partir de leurs fréquences dans le corpus parallèle et

[13] Mots hors vocabulaire

leur phonétisation est automatique. Le ML est un modèle adaptatif de type *cache* décrit au paragraphe 4.5.1, qui utilise le corpus parallèle comme *cache*. Les expérimentations ont été menées sur le corpus TREC-8 SDR couvrant les 6 premiers mois de 1998. Cette méthode réduit le taux d'erreur de 27,5% à 26,7% sur le sous-corpus de 10h dédié à l'évaluation de la reconnaissance, et un gain significatif est observé sur la tâche d'indexation.

Adaptation des ML pour la transcription de documents d'archive

Les expériences précédentes s'appliquent aux documents d'actualités récents dans la mesure où des corpus de textes couvrant la même période sont disponibles. Pour les documents d'archives, tels que le corpus *ECHO* décrit ultérieurement (8.1.1), des textes contemporains ne sont pas forcément disponibles. Ces documents d'archives sont parfois accompagnés d'une courte liste de mots clés en guise de métadonnées. Mais cette liste est trop courte pour estimer une distribution unigramme. Comme aucune autre donnée d'entraînement spécifique au corpus de test n'est disponible, l'article [Federico, 2002] propose d'utiliser un corpus récent de journaux (650M de mots) indexé via une analyse dite *PLSA* (*Probabilistic Latent Semantic Analysis*) afin de générer une distribution unigramme à partir de la liste de mots clefs. Cette distribution permet d'adapter le modèle de langage de référence suivant la méthode *MDI* présentée précédemment.

L'analyse dite *LSA* (*Latent Semantic Analysis*) s'appuie sur une matrice mot-document $A = (N(w_i, d_j))$ qui est une matrice $m \times n$ où m est le nombre de mots et n le nombre de documents. La grandeur $N(w_i, d_j)$ est souvent soit la probabilité unigramme du mot w_i dans le document d_j soit le score $tf * idf$. Dans sa forme standard, la *LSA* calcule les valeurs propres de la matrice A et l'espace sémantique latent (*latent semantic space*) est réduit en choisissant les valeurs propres les plus grandes. La version probabiliste, PLSA, utilise la factorisation en matrice non-négative (*NMF*) où la matrice A est approximée par le produit de de deux matrices non-négatives[Lee and Seung, 2000] :

$$\begin{aligned} A_{ij} = P(w_i|d_j) & \approx \textstyle\sum_{a=1}^{r} P(w_i|t_a)P(t_a|d_j) \\ & \approx \textstyle\sum_{a=1}^{r} W_{ia}H_{aj} = (WH)_{ij} \end{aligned} \tag{4.19}$$

où $r \ll \min(n, m)$ est la dimension réduite. Les matrices W et H sont estimées via l'algorithme EM afin de minimiser la divergence de Kullback-Leibler entre A et le produit WH. Les r colonnes de W représentent une base de r documents. Cette base est alors utilisée pour estimer la distribution unigramme ayant généré la courte description notée d selon le maximum de vraisemblance :

$$\operatorname*{argmax}_{H} \sum_{i=1}^{m} [N(w_i, d) \sum_{a=1}^{r} W_{ia} \cdot H_{a,d}] \tag{4.20}$$

où $N(w_i, d)$ désigne le nombre d'occurrences du terme w_i dans la description d. Le vecteur obtenu est la distribution unigramme $\hat{P}_A(w)$ utilisée pour l'adaptation *MDI* conformément à l'équation 4.17. Cette expérimentation a été menée avec un corpus de journaux de 1,25M de documents et un vocabulaire de 200k mots pour transcrire cinq documents choisis du corpus *ECHO* (une durée totale de 55 minutes). Le modèle de langage de référence est un modèle trigramme standard avec un vocabulaire étendu de 63,5k à 65k mots selon la méthode exposée au paragraphe précédent. Les mots nouveaux ont été sélectionnés en choisissant les plus probables selon la distribution $\hat{P}_A(w)$. Un gain relatif du WER de 3,2% est obtenu.

Conclusion du chapitre

Le rôle d'un modèle de langage lorsqu'il s'inscrit dans un processus d'indexation automatique de documents audiovisuels est central puisqu'il conditionne le contenu linguistique des transcriptions automatiques produites par le système de reconnaissance automatique de la parole. Si dans d'autres domaines d'applications les techniques utilisées en modélisation du langage peuvent varier[14], le consensus est flagrant en ce qui concerne les systèmes de transcription automatique de documents audiovisuels. À titre d'exemple, regardons les trois meilleurs systèmes présentés lors des récentes évaluations des systèmes de transcriptions automatiques organisées par le NIST[15] [Gauvain *et al.*, 2003; Nguyen *et al.*, 2003; Kim *et al.*, 2003]. Les systèmes présentés par le LIMSI et BBN[16] utilisent des modèles quadrigrammes de mots, et celui présenté par l'université de Cambridge, un modèle quadrigramme de classes interpolé avec un modèle quadrigramme de mots. Mais les points communs de ces trois systèmes ne se limitent pas à cela. Et les trois présentations insistent sur les données d'entraînement, leur époque et leur utilisation.

Ce chapitre reprend ces aspects de la modélisation du langage, avec comme objectif de présenter le modèle de langage de référence pour le français utilisé dans les expérimentations de cette thèse, mais aussi d'en montrer les faiblesses, et d'introduire la nécessité et la problématique de l'adaptation des modèles de langage.

Les modèles utilisés dans la suite des travaux sont des modèles ngrammes de mots construits selon les étapes décrites à la section 4.4. En particulier, le modèle de langage de référence résulte de l'interpolation linéaire de trois modèles estimés sur trois corpus. Chaque corpus est normalisé comme cela est décrit au paragraphe 4.4.2 et correspond à un type de texte spécifique comme des transcriptions d'émissions ou d'extraits d'émissions radio et télédiffusées, ou des articles de la presse écrite. L'inconvénient de ce type de modèle est la dépendance aux corpus d'entraînement. Par conséquent, pour qu'un modèle de langage soit efficace, l'adéquation entre le type de langage utilisé dans l'émission à transcrire et dans les corpus d'entraînement est nécessaire. Étant données les quantités de données requises lors de la construction d'un modèle de langage qui sont évoquées au paragraphe 4.4.1, il est manifestement impensable de disposer pour toutes les émissions à transcrire de données adéquates en quantité suffisante pour reprendre du début la construction d'un modèle.

Il est donc nécessaire d'adapter les modèles de langage aux documents à transcrire, rejoignant ainsi les conclusions du NIST à propos des campagnes d'évaluation *SDR* présentées au chapitre 2.

[14]Pour des applications comme la reconnaissance de la parole conversationnelle ou les systèmes de dialogues, différents types de modèles de langage sont utilisés comme les modèles connectionnistes, ou les grammaires hors contextes.

[15]Riched Transcription 2003 Summer Workshop, dont les actes sont disponibles en ligne à l'adresse http ://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/index.htm

[16]www.bbn.com

Les techniques d'adaptation des modèles de langage sont introduites à la section 4.5 en fonction des données qu'elles requièrent. Ce choix est motivé par les objectifs de cette thèse. En effet, les types de données d'adaptation disponibles et leur quantité ne sont pas les mêmes suivant que le document à transcrire est le journal télévisé d'hier soir ou une émission d'archive. L'hypothèse qui sera faite par la suite est qu'il est possible d'acquérir des connaissances *a priori* en utilisant des textes qui sont contemporains de l'émission à transcrire. Les travaux décrits dans cette thèse correspondent et s'inspirent donc des méthodes présentées au paragraphe 4.5.3.

Chapitre 5

Construction automatique du vocabulaire

Le vocabulaire est la liste close des N mots pouvant être reconnus par le système de reconnaissance de la parole. Dans les systèmes de transcription automatique d'émissions télé-radiodiffusées, la taille maximale du vocabulaire N est, en général, fixée à 65 535 mots[1]. Le choix du vocabulaire a un fort impact sur la qualité de la transcription automatique, car tout mot absent du vocabulaire ne peut être reconnu et peut causer des erreurs dans son voisinage. Le vocabulaire est donc construit de manière à obtenir la meilleure couverture lexicale possible sur la tâche considérée, c'est à dire les émissions d'actualité télé-radiodiffusées dans le cadre de ce travail. Les transcriptions de ces émissions semblent alors être le matériau idéal pour en estimer le contenu lexical. Mais ces transcriptions sont disponibles en trop petite quantité. Le risque est alors de construire un vocabulaire spécifique à ce corpus. Appliqué à d'autres émissions, à la fois le vocabulaire du français général et le vocabulaire lié à l'actualité (en particulier, les noms propres liés aux événements relatés et les termes spécifiques qui y sont liés), seront modélisés de manière lacunaire. D'autres sources de textes sont utilisées telles que les journaux qui sont disponibles en grande quantité et permettent de couvrir d'une part, le français général, et d'autre part, l'actualité sur une vaste période. La détermination manuelle du vocabulaire se résume à sélectionner les N formes lexicales les plus fréquentes provenant de plusieurs corpus d'apprentissage afin de maximiser la couverture lexicale sur un corpus de développement. Un algorithme de construction automatique du vocabulaire est proposé, qui interpole un ensemble de modèles unigrammes estimés sur les données d'entraînement, afin de minimiser la perplexité du corpus de développement.

Le cadre d'évaluation est tout d'abord présenté à la section 5.1, rappelant les corpus disponibles et leurs usages suivi des critères d'évaluation. La construction manuelle du vocabulaire de référence est ensuite décrite à la section 5.2. À la section 5.3, la problématique liée à la construction du vocabulaire est présentée puis un algorithme automatique est proposé pour la résoudre.

[1]Cette limite dépend de la langue. Par exemple, pour l'allemand, il n'est pas possible de se contenter de 65 535 mots.

Les expérimentations relatées dans la dernière section 5.4 permettent d'évaluer cet algorithme avec différents corpus d'entraînement.

5.1 Cadre d'évaluation

Le vocabulaire est l'ensemble fini des mots connus du système de RAP, et doit par conséquent avoir la meilleure couverture lexicale sur la tâche. L'enjeu de l'évaluation est donc de constituer un ou des corpus qui modélisent la tâche afin d'y mesurer la couverture lexicale. La couverture lexicale d'un vocabulaire sur un texte donné est le ratio entre le nombre d'occurrences des mots du vocabulaire dans le texte, et le nombre total d'occurrences dans le texte. Le dual de la couverture lexicale est le taux de mots hors vocabulaire (MHV) sur un texte donné, qui est le ratio entre le nombre d'occurrences dans le texte de mots absents du vocabulaire, et le nombre total d'occurrences dans le texte. Ces deux quantités somment à un.

Le cadre d'évaluation a pour objectif de comparer le vocabulaire de référence construit manuellement avec des vocabulaires obtenus à l'aide de l'algorithme de construction automatique présenté à la section 5.3. Pour cela, les corpus d'entraînement et le corpus de développement décrits aux paragraphes 5.1.1 et 5.1.2 sont communs aux deux types de vocabulaire.

5.1.1 Corpus d'entraînement

Les corpus utilisés sont les mêmes que ceux décrits au paragraphe 4.4.1. Les textes disponibles se répartissent selon 5 sources. La première contient les transcriptions fines de journaux radio-télévisés qui ont été utilisées lors de l'apprentissage des modèles acoustiques. Ces émissions proviennent de 6 diffuseurs et se répartissent sur 6 années de 1994 à 1999. La deuxième source contient des transcriptions d'extraits d'émission datant des années 1997, 1998 et 2000, réalisées par une agence de service de presse. Les trois dernières sources proviennent de la presse écrite. Ce sont des dépêches d'agences de presse datant de 1994 à 1996, des articles du journal *Le Monde* de 1987 à 1998 et des articles du journal *Le Monde Diplomatique* de 1990 à 1996. Un sous-corpus par année a été constitué pour chacune de ces sources. Le tableau 5.1 donne les nombres moyen, minimum et maximum de mots, pour chaque source calculés par année, ainsi que la période couverte. La dernière colonne totalise le nombre de mots par source. Les transcriptions représentent une faible quantité de mots, justifiant ainsi l'utilisation d'autres sources palliatives.

5.1.2 Critères d'évaluation

Le premier critère d'évaluation du vocabulaire est le taux de MHV sur le corpus de développement. Ce critère caractérise l'aptitude de la méthode à générer un vocabulaire « pour le corpus de développement » sous certaines contraintes détaillées à la section 5.2.2. Le corpus de développement choisi contient 17 026 occurrences pour 7 247 formes et contient des transcriptions manuelles du journal télévisé *6 minutes* de M6 datant de l'année 2000. Ce corpus fut constitué lors de la construction du vocabulaire de référence avec les transcriptions qui étaient alors les

Source	*Moyenne*	*Minimum*	*Maximum*	*Période*	*Total*
Transcriptions	267 302	5003	677 118	1994-1999	1 603 800
Service de presse	24 217 930	2 102 667	49 901 588	1997,1998,2000	72 653 790
Agence de presse	22 271 693	16 475 195	32 382 159	1994-1996	66 815 080
Le Monde	21 450 500	14 993 766	24 522 993	1987-1998	257 406 060
Le Monde Diplo.	959 199	730 474	1 097 475	1990-1996	6 714 390

TAB. 5.1 – Chaque source est caractérisée par le nombre moyen, minimum et maximum de mots par année, ainsi que par la période couverte et le nombre total de mots.

plus récentes. Afin d'évaluer l'aptitude à construire un vocabulaire générique, un corpus de test est choisi. Il contient 77 541 occurrences de transcriptions manuelles du journal télévisé de 20h du mois de janvier 2002 d'une des principales chaînes de télévision française. Ainsi ce corpus est constitué à partir de transcriptions de documents qui sont au moins de 2 ans plus récents que les textes d'entraînement ou de développement. Ces deux critères sont étalonnés par les performances de deux vocabulaires particuliers : le vocabulaire de référence décrit à la section 5.2 et le vocabulaire idéal présenté dans le paragraphe suivant.

5.1.3 Le vocabulaire idéal

Si la taille du vocabulaire n'est pas limitée, on peut envisager de prendre toutes les formes apparaissant dans tous les corpus d'entraînement. Dans notre cas, ce vocabulaire idéal contient 765 442 formes distinctes. La couverture lexicale obtenue avec ce vocabulaire, que ce soit sur le corpus de développement ou sur le corpus de test, est bien entendu la meilleure possible pour l'ensemble de corpus d'entraînement donné. Ainsi, la borne inférieure du taux de MHV est de 0,25% sur le corpus de développement et de 0,35% sur le corpus de test. Le nombre de MHV sur le corpus de développement est de 42 occurrences pour 32 formes qui sont données dans le tableau 5.2 avec leurs fréquences d'apparition.

Ces mots sont, en grande majorité, des entités nommées : 26 sur 32, soit un peu plus de 80%. Parmi les entités nommées, les formes composées (« Total-Fina », « Saint-Bart », « Hachette-Filipacchi », « Auto-Journal » et « Aix-Marseille »), sont des formes classées hors vocabulaire à cause de la normalisation des mots composés. Ceci illustre le propos du paragraphe 4.4.2 sur la difficulté à faire des choix lors de la normalisation. L'exemple « Total-Fina » est à ce titre intéressant car suivant l'époque du texte, il est possible de rencontrer d'une part « Total », « Fina » et « Elf », et d'autre part, « Total-Fina » (en 2000) puis « Total-Fina-Elf » (en 2001/2002, plus récemment la forme TotalFinaElf peut être rencontrée). Un même texte peut aussi contenir la forme « Total-Fina-Elf » pour faire référence au groupe puis simplement « Elf » pour aborder le procès du même nom. Sur les 6 formes qui ne sont pas des entités nommées, deux sont des adjectifs dont un (« tectiques ») est inconnu du dictionnaire « les trésors de la langue française »[2] et du moteur de recherche Google[3], et l'autre (« ridiculisante ») est une

[2] http ://atilf.inalf.fr/tlfv3.htm
[3] www.google.com

Rézala	4	Provil	1
Organiz	4	Ponsol	1
Larusso	2	PACSant	1
Fleutiot	2	Obouchi	1
D2	2	Monchoi	1
Cinésix	2	Hachette-Filipacchi	1
Walert	1	Gonaïve	1
Vilcoc	1	fatalitis	1
Varancheville	1	Eblemguem	1
Total-Fina	1	D1	1
tectiques	1	Blarengem	1
Tarone	1	Bénélucky	1
Sandakane	1	BCEM	1
Saint-Bart	1	Auto-Journal	1
ridiculisante	1	Angien	1
Quess	1	Aix-Marseille	1

TAB. 5.2 – Les MHV du vocabulaire idéal obtenus sur le corpus de développement avec leurs nombre d'occurrences.

inflexion rare de son verbe. Il reste deux sigles sportifs (« D1 » et « D2 »), un mot étranger (« fatalitis ») et un participe présent dérivé d'un acronyme (« PACSant »). Les MHV ne sont donc que des mots rares et seules 6 formes apparaissent plus d'une fois. A l'exception de « D2 » et de « Cinésix » (une émission de télévision), les formes apparaissant plus d'une fois désignent des personnes ou un groupe de personnes ayant eu une importance et une existence très ponctuelle dans l'actualité.

Les performances du vocabulaire idéal montre qu'en utilisant de manière optimale les ressources, il est possible d'atteindre un taux de MHV de 0,25%. Les MHV sont principalement des entités nommées qui apparaissent très ponctuellement dans l'actualité.

5.2 Le vocabulaire de référence

Le vocabulaire de référence a été construit en 2000, à partir des textes décrits au paragraphe 5.1.1 et selon la procédure manuelle détaillée dans le paragraphe 5.2.1. Il s'agit d'une réactualisation des travaux relatés dans les articles [Adda-Decker *et al.*, 1999; Adda *et al.*, 1997b] en fonction de nouvelles données. Le cadre théorique dans lequel s'inscrit la construction du vocabulaire est d'abord décrit, permettant d'en dégager les contraintes pratiques. Puis le processus de construction est présenté, puis le vocabulaire de référence est ensuite évalué et une typologie des mots hors vocabulaire est effectuée.

5.2.1 Cadre théorique

Le vocabulaire est construit de manière à obtenir la meilleure couverture lexicale possible sur la tâche. Son contenu lexical ne peut être connu directement, même si l'on sait que dans notre cas, qu'il s'agit d'émissions d'actualité. Un corpus de développement est alors constitué à partir de transcriptions de journaux télévisés. À partir de ce corpus, une distribution unigramme sur les mots p_Y est estimée qui permet de modéliser le contenu lexical de la tâche. L'objectif est alors choisir le vocabulaire V parmi l'ensemble des vocabulaires possibles contenant des mots du corpus d'entraînement, qui minimise l'espérance du taux de mots hors vocabulaire T_{MHV} connaissant la distribution p_Y, ce qui s'écrit :

$$V = \operatorname*{argmin}_v E[T_{MHV}(v)|p_Y] \tag{5.1}$$

Pour estimer l'espérance $E[T_{MHV}(v)|p_Y]$, la fonction δ_v est définie qui indique si un mot w est hors vocabulaire. L'espérance s'écrit alors de la manière suivante :

$$\begin{cases} E[T_{MHV}(v)|p_Y] = \sum_w \delta_v(w) p_Y(w), \\ \text{avec } \delta_v(w) = 1 \text{ si } w \notin v \\ \text{et} \quad \delta_v(w) = 0 \text{ sinon.} \end{cases} \tag{5.2}$$

Une contrainte importante pour un vocabulaire est sa taille N. Si le vocabulaire est vide, tous les mots sont hors vocabulaire et l'espérance du taux de MHV est maximum : $E[T_{MHV}(v)|p_Y] = \sum_w p_Y(w) = 1$. Si maintenant, il est possible de construire un vocabulaire contenant un seul mot ($N = 1$), ce mot est choisi tel que $E[T_{MHV}(v)|p_Y]$ soit minimum. Il s'agit du mot le plus probable. Si la taille du vocabulaire augmente à nouveau ($N = 2$), le second mot sera le second le plus probable. Ainsi, le vocabulaire de N mots qui minimise l'espérance du taux de mots hors vocabulaire connaissant p_Y, est celui qui contient les N mots les plus probables.

Mais il n'existe pas assez de transcriptions pour constituer un corpus de développement permettant la construction d'un vocabulaire de $N = 65535$ mots. Dans notre cas, le corpus de développement permet au mieux la sélection de 7 247 mots. Mais si cela est possible, le danger est de construire un vocabulaire spécifique au corpus de développement qui est une modélisation lacunaire de la tâche, en particulier en ce qui concerne le vocabulaire du français général. Il est donc préférable que le corpus de développement serve uniquement pour calculer le taux de MHV, et que des corpus d'entraînement soit utilisés pour la sélection des mots.

5.2.2 Construction du vocabulaire de référence

Construire le vocabulaire se résume à sélectionner les N formes lexicales les plus fréquentes provenant de plusieurs corpus d'apprentissage afin de maximiser la couverture lexicale sur un corpus de développement. L'énoncé est simple, mais la pratique se révèle être plus délicate. L'hypothèse de caractériser un mot par sa fréquence d'apparition est implicitement faite au paragraphe 5.2.1. L'hypothèse est forte, car la fréquence est une grandeur discrète et dépend du corpus sur lequel elle est calculée. Deux conséquences immédiates sont à prendre en compte.

D'abord, la taille des corpus d'entraînement doit être suffisante pour permettre la construction d'un vocabulaire suffisamment générique pour être appliqué à des tâches autres que celle représentée par le corpus de développement. Dans le cas où il n'y a qu'un corpus d'apprentissage, il suffit de sélectionner les N formes les plus fréquentes. Puis, l'utilisation de seuils sur le rang pose le problème des *ex aequo* : plusieurs mots peuvent avoir la même fréquence. Pour éviter tout choix aléatoire, tous les mots ayant la même fréquence sont sélectionnés. Le seuillage sur les fréquences ne permet donc pas de sélectionner N mais M mots, avec $M \leq N$.

En pratique, plusieurs corpus d'apprentissage sont utilisés. Comme le montre le tableau 5.1, il est préférable de ne pas regrouper toutes les données dans un seul corpus. Sinon, étant donné les différences de taille, les transcriptions manuelles peuvent être noyées dans les millions de mots provenant des autres corpus.

Pour chaque corpus, un seuil sur la fréquence des mots élimine les mots les moins fréquents. Une liste de mots par corpus est alors obtenue, et le vocabulaire final est la fusion de ces listes. Chaque seuil est le résultat d'un compromis entre trois critères : le nombre de mots N que doit contenir le vocabulaire final, la couverture lexicale sur le corpus de développement et le filtrage de bruit éventuel. Les deux premiers critères concernent le vocabulaire final et ne peuvent être connus qu'après sa construction, alors que le dernier est propre à chaque source. Le choix des seuils est donc crucial. Cette procédure est manuelle car les trois critères ne suffisent pas à fixer un ensemble unique de seuils et d'autres considérations peu quantifiables interviennent dans le choix des seuils, telles que la généricité, la taille ou la nature des sources. Si un nombre important de sources est disponible comme c'est le cas ici, il est impératif d'effectuer des regroupements, car il est difficile de construire manuellement un vocabulaire avec un nombre important de sources. Il a été choisi de regrouper les textes selon leur nature en ne tenant pas compte de leur date. Trois corpus ont été ainsi constitués : le premier regroupe tous les textes de la presse écrite (*Le Monde*, *Le Monde diplomatique*, Agence de presse), le deuxième les textes de services de presse et le troisième les transcriptions fines de journaux radio-télévisés.

Le vocabulaire de référence ainsi construit contient 65 333 formes. Son taux de MHV sur le corpus de développement est de 0,91%[4] ce qui représente 155 occurrences pour 132 formes. Sur le corpus de test, les performances sont un peu dégradées avec un taux de MHV de 1,13%, soit 873 occurences pour 604 formes. Les formes hors vocabulaire diffèrent, et seules 2 formes hors vocabulaire sont communes aux deux corpus : « tombola » et « coupe-feu ». En deux ans, la population des MHV a donc changé quasiment du tout au tout et ce sont en grande majorité des mots qui apparaissent une seule fois dans le corpus de test.

5.2.3 Typologie des mots hors vocabulaire

Les corpus de développement et de test ont été étiquetés en catégories morpho-syntaxiques (*part of speech*), à l'aide du tagger de Brill[5] [Brill, 1994] adapté au français par l'INALF[6]. Le jeu

[4]Le taux de MHV était de 1,2% lors de sa construction. Mais la normalisation du corpus de développement a été modifiée dans le cadre de ces travaux afin de l'homogénéiser avec celle du corpus de test.

[5]Disponible à http ://www.cs.jhu.edu/~brill

[6]Disponible à http ://jupiter.inalf.cnrs.fr/WinBrill

d'étiquettes a été réduit après l'opération de marquage, en supprimant l'information de nombre. La répartition des MHV par catégorie morpho-syntaxique est donnée dans le tableau 5.3. Plus de la moitié des MHV sont des noms propres et presque 15% sont des noms communs. Peu de différences sont observées dans la répartition des MHV entre le corpus de développement et le corpus de test.

	% des MHV	
Marqueur	Dev	Test
Nom Propre	50,3	55,4
Nom commun	14,2	14,3
Verbe conjugué	8,4	13,1
Adjectif	8,4	4,1
Verbe infinitif	5,2	3,1
Participe passé adjectif	4,5	4,3
Participe passé (être)	3,9	2,2
Participe passé (avoir)	1,9	1,4
Mot étranger	1,3	0,1
Participe présent	1,3	0,9
Adverbe	0,6	0,7

TAB. 5.3 – Répartition des marqueurs morpho-syntaxiques des MHV du vocabulaire de référence observés sur le corpus de développement

5.3 Construction automatique du vocabulaire

L'objectif est de construire un vocabulaire de N mots à partir de K corpus d'entraînement notés $(X_1, X_2, \ldots, X_K)$, afin d'optimiser la couverture lexicale sur un corpus de développement Y. Selon le cadre théorique présenté au paragraphe 5.2.1, le corpus de développement n'est pas accessible pour la sélection des mots, la combinaison des corpus d'entraînement doit donc permettre d'approcher la distribution observée sur le corpus de développement. La construction du vocabulaire nécessite de choisir un opérateur de mélange, et de choisir une fonction coût.

5.3.1 Segmentation des corpus d'entraînement

Une représentation statistique d'un corpus est sa distribution unigramme. Beaucoup d'informations sont alors perdues, mais cette modélisation s'avère satisfaisante pour certaines tâches, en particulier pour la recherche d'information et la détection de thème. Par la suite, une distribution unigramme de mots estimée sur le corpus X sera notée : p_X. Cette distribution attribue la probabilité $p_X(w)$ au mot w.

Lors de la construction du vocabulaire de référence décrite à la section 5.2, les textes ont été regroupés selon le type de source, sans prendre en considération leurs dates. L'usage des trois sources[7] se justifie par le fait que peu de données représentatives de la tâche (les transcriptions manuelles) sont disponibles. Un vocabulaire construit avec si peu de données risque d'être très incomplet et de mal représenter la composante générale du langage. Le manque de données s'illustre par les 36 000 formes distinctes observées sur l'ensemble des transcriptions manuelles, ce qui est peu en comparaison avec la limite du vocabulaire de 65 535 mots. La deuxième source contient des transcriptions d'extraits de documents d'actualité télé-radiodiffusés. Représentatif de la tâche, ce corpus est 7 fois plus important que le corpus de transcriptions manuelles. Mais ce sont les journaux qui assurent une quantité de données suffisante et donc une meilleure modélisation du français générique.

La segmentation des données d'entraînement en plusieurs corpus est indispensable pour conserver la spécificité de chacun. Si par exemple, les textes de la presse écrite et les transcriptions manuelles sont mélangés, étant donné qu'il y a presque un rapport de taille de 200 en faveur des textes de la presse écrite, même les mots les plus fréquents dans les transcriptions manuelles deviennent peu fréquents dans le mélange. Dans une mise en commun des textes, les transcriptions manuelles sont occultées par les textes de la presse écrite.

Cependant, dans le cas de la construction manuelle du vocabulaire, cette segmentation est limitée à trois corpus. Dans la perspective d'une construction automatique, cette limite peut être remise en cause. À la section 5.4, plusieurs configurations seront décrites et expérimentées en ce qui concerne la séparation des corpus d'entraînement.

5.3.2 Choix des vocabulaires pour l'estimation des distributions

Pour permettre la combinaison de différents modèles unigrammes, deux solutions peuvent être envisagées. La première est d'utiliser un grand vocabulaire commun regroupant toutes les formes apparaissant dans tous les corpus d'apprentissage. Ainsi, au problème des mots hors vocabulaire se substitue celui des *zerotons*, qui sont, pour un modèle donné, les mots du vocabulaire qui ne sont pas observés sur son corpus d'entraînement. Pour attribuer une probabilité non nulle aux *zerotons*, une masse de probabilité non négligeable est prélevée sur les mots observés lors de l'estimation (pour plus de détails, se reporter au paragraphe 4.4.4).

La seconde solution est d'estimer les distributions indépendemment l'une de l'autre avec un vocabulaire qui leur est propre. Dans ce cas pour assurer la compatibilité des vocabulaires entre eux, la forme <UNK> est intégrée à tous les vocabulaires et représente les mots hors vocabulaire. Il est d'usage lors de l'entraînement d'un modèle de langage de projeter toutes les formes hors vocabulaire sur la forme <UNK>. La forme <UNK> a alors une fréquence d'apparition et une probabilité relativement importante dans le corpus d'entraînement. Dans notre cas, le choix a été fait de ne pas avoir de formes hors vocabulaire lors de l'entraînement[8]. Par défaut, la probabilité quasi-nulle de 10^{-99} est attribuée à la forme <UNK>.

[7]Les trois sources sont des transcriptions d'émissions d'actualité, transcriptions d'extraits d'émissions , et des textes de la presse écrite.

[8]Pour chaque modèle, le vocabulaire est composé de l'ensemble des formes rencontrées dans le corpus d'entraînement et de <UNK>.

En pratique la deuxième solution apparaît comme plus satisfaisante car dans la première solution la probabilité attribuée à un zéroton est dépendante du corpus d'entraînement. Dans la deuxième solution tous les zérotons sont projetés sur la forme <UNK> dont on fixe au préalable la probabilité. Lors des expérimentations relatées au paragraphe 5.4, la différence entre ces deux méthodes ne s'est pas montrée pertinente. Cependant, la méthode utilisant un vocabulaire propre à chaque corpus semble être plus performante et c'est celle qui sera adoptée par la suite. L'estimation des probabilités est alors effectuée selon le principe de maximum de vraisemblance.

5.3.3 Opérateur et fonction coût

Le mélange des distributions d'entraînement est réalisé par interpolation linéaire. L'objectif initial peut alors être reformulé ainsi : calculer le jeu de coefficients d'interpolation

$$\Lambda = (\lambda_1, \lambda_2, \dots \lambda_K), \text{ avec } \sum_{i=1}^{K} \lambda_i = 1,$$

qui minimisent une fonction de coût $\mathcal{C}$ entre la distribution estimée sur le corpus de développement p_Y et la distribution interpolée $\tilde{p}_X$. Ce qui s'écrit de la manière suivante :

$$\begin{aligned} \Lambda^* &= \underset{\Lambda}{\operatorname{argmin}}\, \mathcal{C}(p_Y, \tilde{p}_X) \\ &= \underset{\Lambda}{\operatorname{argmin}}\, \mathcal{C}(p_Y, \sum_{i=1}^{K} \lambda_i \; p_{X_i}), \end{aligned} \tag{5.3}$$

Un mot étant représenté par sa probabilité, la construction du vocabulaire à partir de la distribution $\tilde{p}_X$ se fait en deux temps : elle nécessite le classement des mots par ordre décroissant de probabilité et la sélection des N mots les plus probables. Cette sélection se fait donc par seuillage sur le rang lexical des mots et non sur leurs fréquences. Ceci implique que le vocabulaire et par conséquent la couverture lexicale ne sont pas directement accessibles.

La fonction de coût est une fonction des deux distributions p_Y et $\tilde{p}_X$ que l'on suppose corrélée avec la couverture lexicale du vocabulaire obtenu à partir de $\tilde{p}_X$. Si le corpus de développement est représentatif de la tâche, sa distribution est celle attendue dans les documents qui vont être traités. Les coefficients d'interpolation peuvent être calculés afin de minimiser le coût d'émettre comme hypothèse $\tilde{p_X}$ alors que la distribution réelle est p_Y. En théorie de l'information, la divergence de Kullback-Leibler quantifie le coût de cette hypothèse. L'équation 5.3 s'écrit alors :

$$\begin{aligned} \Lambda^* &= \underset{\Lambda}{\operatorname{argmin}} \sum_{w \in \mathcal{V}} p_Y(w) \log \left(\frac{p_Y(w)}{\sum_{i=1}^{K} \lambda_i \; p_{X_i}} \right) \\ \Lambda^* &= \underset{\Lambda}{\operatorname{argmax}} \sum_{w \in \mathcal{V}} p_Y(w) \log (\sum_{i=1}^{K} \lambda_i \; p_{X_i}) \end{aligned} \tag{5.4}$$

Puisque les modèles utilisés sont des modèles unigrammes, cette approche est équivalente à celle couramment rencontrée pour calculer les coefficients d'interpolation : minimiser la perplexité. Cependant, cette fonction coût n'est pas la couverture lexicale, et l'équivalence supposée entre la perplexité et la couverture lexicale sera étudiée au paragraphe 5.4.4.

5.3.4 Algorithme de construction automatique du vocabulaire

À partir de chacun des corpus d'entraînement, un modèle unigramme est estimé. Ces distributions sont ensuite interpolées et l'algorithme E.M [Dempster *et al.*, 1977] est utilisé pour calculer le jeu de coefficients d'interpolation qui minimise la perplexité du corpus de développement. Le vocabulaire est construit à partir de cette distribution interpolée en choisissant les N mots les plus fréquents. Le nombre N (65 333) est le même que la taille du vocabulaire de référence permettant ainsi de comparer les vocabulaires à taille égale. L'algorithme peut alors se décrire de la manière suivante.

1. Estimation à partir de chaque corpus d'entraînement X_i d'une distribution unigramme p_{X_i} pour $i = 1$ à K.
2. Initialisation des coefficients d'interpolation, $(\lambda_i = 1/K)$.
3. Application de l'algorithme E.M pour calculer les coefficients d'interpolation afin de minimiser la perplexité du corpus de développement.
4. Construction du vocabulaire à partir de la distribution interpolée $\tilde{p}_X$, en sélectionnant les $N = 65\ 333$ mots les plus fréquents

Dans chaque distribution, les mots du corpus de développement qui ne figurent pas dans le vocabulaire de X_i sont considérés comme hors vocabulaire et leurs probabilités sont celles de <UNK> (voir paragraphe 5.3.2). Les mots du corpus de développement qui n'appartiennent à aucun des corpus d'entraînement ne sont pas pris en compte. En ce qui concerne l'algorithme E.M, la condition d'arrêt a été fixée en terme de précision sur les coefficients d'interpolation. Cette précision dépend du nombre de corpus d'entraînement utilisés.

5.4 Expérimentations

L'algorithme décrit précédemment permet de manipuler autant de corpus que nécessaire. À partir des corpus décrits au paragraphe 5.1.1, plusieurs segmentations peuvent être envisagées. Les quatre segmentations qui nous ont semblé être les plus significatives sont :
- la segmentation *originale* est celle utilisée pour la construction du corpus de référence,
- la segmentation par *année et source* sépare les données selon leur source et l'année,
- la segmentation *hybride* sépare les données par diffuseurs pour les transcriptions et par années pour les autres sources,
- la segmentation *année et type de source* sépare les textes selon le type de leurs sources (presse écrite, transcription extraits d'émissions, et transcription) et par année.

	Dev 2000		Test 2002	
Vocabulaire	%MHV	#MHV	%MHV	#MHV
Standard	0,91	155	1,13	873
Idéal (765 442 mots)	0,25	42	0,35	274
Original	0,79	134	1,06	822
Années et sources	0,79	135	1,05	814
Années et sources hybride	0,78	132	1,05	819
Années et type de source	0,76	129	1,05	818

TAB. 5.4 – Évaluation des choix de segmentation de corpus en terme de taux de MHV mesurés sur le corpus de développement « Dev 2000 » et sur le corpus de test « Test 2002 »

Pour chaque segmentation un vocabulaire est construit puis évalué. Le tableau 5.4 résume l'ensemble des résultats en évaluant chaque vocabulaire par son nombre et son taux de MHV sur le corpus de développement (Dev 2000) et sur le corpus de test (Test 2002). Les performances du vocabulaire de référence et du meilleur vocabulaire y sont également rappelées.

5.4.1 Validation

La première expérimentation vise à valider l'algorithme décrit à la section précédente en l'utilisant dans les mêmes conditions que celle de la construction du vocabulaire de référence, avec les mêmes corpus d'entraînement. Les corpus d'entraînement sont regroupés selon la nature des textes : transcription d'émissions radio-télédiffusées, transcription d'extraits d'émissions et presse écrite. Le vocabulaire construit automatiquement avec la même segmentation des données d'entraînement est dénommé « Original » dans le tableau 5.4 qui résume l'ensemble des résultats. La méthode de construction automatique améliore nettement le vocabulaire à corpus identiques, que ce soit sur le corpus de développement avec un gain relatif de 12%, ou sur le corpus de test dans une moindre mesure, avec un gain relatif est de 6%.

5.4.2 Segmentation par années et par sources

La segmentation par années est la segmentation expérimentée la plus fine : pour chaque année de chaque source un modèle est construit, obtenant ainsi 30 modèles. Sur le corpus de développement, le taux de MHV observé avec ce vocabulaire est équivalent à celui observé avec le vocabulaire « original ». Sur le corpus de test, le taux de MHV est sensiblement réduit.

Les performances obtenues avec ces deux vocabulaires construits automatiquement sont donc comparables. L'examen des coefficients d'interpolation permet de quantifier la participation de chacun des corpus au vocabulaire. Le tableau 5.5 présente les corpus ayant une contribution non négligeable et leur coefficients d'interpolation. Les corpus construits sur les textes du journal *Le Monde diplomatique* ne semblent pas du tout adaptés puisque leurs coefficients d'interpolations sont nuls. Pour d'autres corpus, les coefficients sont également nuls ou négligeables. Sur l'ensemble du tableau, la corrélation entre l'année du corpus et l'importance de sa contribution n'est pas

Coefficient	Source	Année
0,8070	Service de presse	2000
0,0526	Transcription	1998
0,0347	*Le Monde*	1998
0,0333	Agence de presse	1995
0,0179	Agence de presse	1996
0,0165	*Le Monde*	1997
0,0147	*Le Monde*	1996
0,0084	*Le Monde*	1995
0,0072	Transcription	1999
0,0040	*Le Monde*	1994
0,0034	*Le Monde*	1992
0,0002	Service de presse	1997

TAB. 5.5 – Coefficients d'interpolation calculés via l'algorithme E.M et leurs corpus d'entraînement associés

manifeste. Cependant, en analysant le tableau par source, l'impact du temps se révèle. Pour *Le Monde*, les corpus sont rangés dans l'ordre chronologique à l'exception de l'année 1993, en partant du plus récent et plus proche du corpus de développement (1998) pour s'arrêter à 1992. Pour les transcriptions et les textes d'agences de presse, les deux années les plus récentes sont retenues. Le corpus « Service de presse » 2000 est tellement prépondérant qu'il fait figure d'exception. L'algorithme E.M. lui attribue un coefficient de l'ordre de 0,8 alors que le second corpus à un coefficient de 0,05. Le fait que ce soit le seul corpus de la même année que celui de développement n'y est sans doute pas étranger.

5.4.3 Segmentation hybride

Une variante de cette segmentation consiste à regrouper les transcriptions non plus par année, mais par diffuseurs. Le nombre de modèles reste inchangé et leurs tailles sont comparables. Cette configuration est notée « Années et sources hybride » dans le tableau 5.4. Elle améliore sensiblement le taux de MHV sur le corpus de développement. L'étude des coefficients montre cependant d'importants changements dans les contributions de chaque corpus. Si le corpus « Service de presse 2000 » reste prépondérant avec un coefficient de 0,79, désormais 14 corpus au lieu de 11 ont une importance non négligeable. En particulier, les transcriptions occupent une part plus importante avec 5 corpus au lieu de 2. Les textes du journal *Le Monde diplomatique* sont exclus de cette variante comme de la précédente.

Le meilleur taux de MHV est obtenu en regroupant tous les textes de la presse écrite par année, les extraits d'émissions par année également et les transcriptions par diffuseur comme précédemment. Ce vocabulaire est référencé « Années et type de source ».

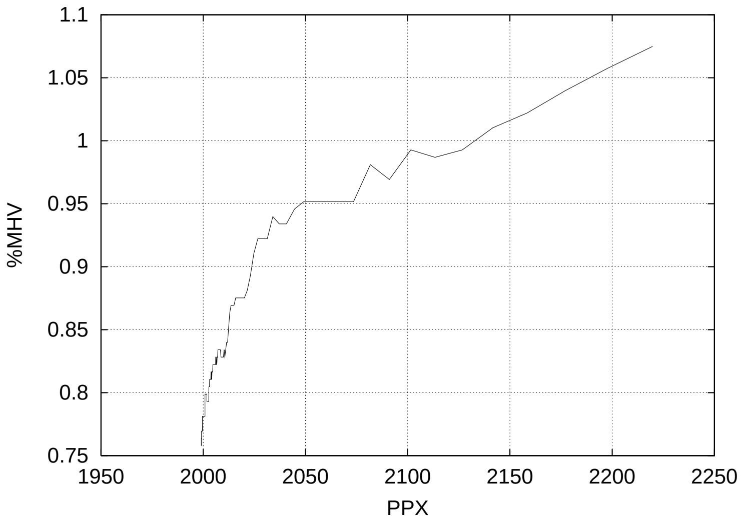


FIG. 5.1 – Taux de MHV en fonction de la perplexité du corpus de développement pour la segmentation *Années et type de source* obtenus lors des itérations de l'algorithme E.M.

5.4.4 Équivalence entre perplexité et couverture lexicale

Au paragraphe 5.3.3, l'hypothèse est faite que minimiser la perplexité calculée avec un modèle interpolé, implique la minimisation du taux de MHV du vocabulaire construit à partir de ce modèle. L'algorithme E.M. est un algorithme itératif qui à chaque itération calcule les coefficients d'interpolation et réduit la perplexité. La figure 5.1 trace le taux de MHV obtenu sur le corpus de développement en fonction de la perplexité pour les différentes itérations de l'algorithme E.M. La segmentation des corpus d'entraînement correspond à la configuration « Années et type de source ». Les corpus ayant une contribution négligeable ont été retirés permettant une convergence plus rapide de l'algorithme.

Nous avions implicitement supposé la fonction monotone croissante, ce qui est globalement le cas. Pour mieux comprendre les petites variations de la courbe, la figure 5.2 représente l'évolution du nombre de MHV en fonction des 200 premières itérations de l'algorithme E.M. Cette fois-ci la fonction était supposée monotone décroissante mais quelques augmentations sont pourtant observées. Cependant, ces augmentations du nombre de MHV ont toujours pour amplitude 1 mot à une exception près (de 2 mots). Elles sont par ailleurs relativement rares puisqu'elles n'interviennent que 8 fois sur les 1 520 itérations (24 modèles sont interpolés). Ces irrégularités peuvent donc être imputées à la sélection des mots du vocabulaire par seuillage sur le rang lexical. Au delà de l'itération 200, la fonction est effectivement monotone décroissante.

Conclusion du chapitre

Le vocabulaire est une base primordiale des connaissances du système de reconnaissance automatique de la parole sur le langage. Phonétisé, il devient le dictionnaire de prononciation, et

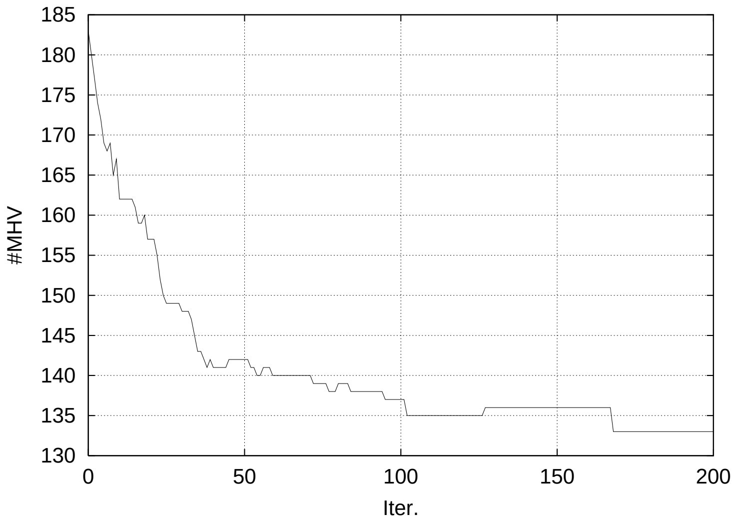


FIG. 5.2 – Nombre de MHV en fonction des itérations de l'algorithme E.M.pour la segmentation *Années et type de source* (24 modèles sont interpolés)

le modèle de langage est construit de manière à attribuer une probabilité à toutes les séquences de mots qui en sont extraites. Dans le cadre de la transcription automatique de documents audiovisuels, la couverture lexicale du vocabulaire sur le document à transcrire est déterminante. En effet, tout mot prononcé dans le document qui n'est pas dans le vocabulaire génère des erreurs de reconnaissance. Et dans le cadre d'un système d'indexation automatique, tout mot hors vocabulaire ne peut être reconnu, ce qui peut créer une lacune au niveau des recherches ultérieures dans la transcription.

Le vocabulaire de référence est construit afin d'obtenir la meilleure couverture lexicale possible sur la tâche de transcription d'émissions d'actualité télé-radiodiffusées. Un corpus de développement est choisi pour représenter cette tâche. Il est constitué de transcriptions de journaux télévisés datant de l'année 2000. La construction du vocabulaire est en partie similaire à celle du modèle de langage. En particulier, les données d'entraînement sont les mêmes, et elles sont regroupées de la même façon selon le type de textes (presse écrite, transcription d'émissions ou d'extraits d'émissions) et couvrent les années 1988 à 2000. Le vocabulaire est évalué par son taux de mots hors vocabulaire observé sur le corpus de développement et sur un corpus de test constitué, lui aussi, de transcriptions de journaux télévisés. Le corpus de test date du mois de janvier 2002 et se distingue du reste des données par son époque qui est de deux ans plus récente.

Une typologie des mots hors vocabulaire observés sur le corpus de développement et le corpus de test montre que plus de la moitié d'entre eux sont des noms propres ; cette part augmente de 10% en relatif en deux ans (de 2000 à 2002). Ceci confirme la nécessité de réduire le taux de mots hors vocabulaire dans une perspective d'indexation. En effet les mots hors vocabulaire concernent principalement les noms propres et cette catégorie est corrélée aux performances d'indexation automatique de type *SDR* (voir paragraphe 2.2.5).

Le processus manuel de construction du vocabulaire est décomposé afin de mettre en place un algorithme de construction automatique du vocabulaire. Cet algorithme consiste en l'interpolation d'un ensemble de modèles unigrammes estimés sur les données d'entraînement, les coefficients d'interpolation sont calculés grâce à l'algorithme E.M. afin de minimiser la perplexité du corpus de développement. Le nombre de corpus étant, en théorie, illimité, différentes segmentations des données d'entraînement sont expérimentés. Globalement, tous les modèles construits automatiquement ont des meilleures performances que le vocabulaire de référence en terme de taux de mots hors vocabulaire, avec dans le meilleur des cas un gain relatif de 17% sur le corpus de développement et 6,3% sur le corpus de test. Cet algorithme repose sur l'hypothèse que minimiser la perplexité du corpus implique une amélioration de la couverture lexicale. Cette hypothèse est vérifiée expérimentalement, validant ainsi cette approche.

Chapitre 6

Adaptation du vocabulaire

La plupart des systèmes de transcription automatique de documents d'actualité utilisent un vocabulaire et un modèle de langage développés à partir de corpus d'apprentissage figés. Le coût lié à la préparation des corpus et les quantités de données nécessaires motivent en partie ce choix (voir la description des corpus et de leur préparation à la section 5.2). Mais le défaut de cette approche est que les données d'apprentissage ont un décalage temporel avec le document à transcrire (un an est courant). Or dans un cadre d'application comme la transcription automatique de documents d'actualité, cela n'est pas sans poser problème. En effet, les caractéristiques linguistiques des documents à traiter dépendent fortement de l'actualité. En particulier, le taux de mots hors vocabulaire (MHV) augmente lorsque le document à traiter et le corpus d'entraînement sont éloignés dans le temps. En moyenne, chaque mot hors vocabulaire est la cause de 1,5 à 2 erreurs. Représentant environ la moitié des MHV, les entités nommées sont caractérisées par des distributions particulières qui changent avec l'actualité [Béchet and Yvon, 2000]. Elles ont une incidence certaine sur la qualité de l'indexation [Garofolo *et al.*, 1998].

Une solution consiste à réactualiser le corpus d'apprentissage puis à ré-estimer le vocabulaire et le modèle de langage. Mais encore faut-il pouvoir accéder à un corpus récent et surtout de taille suffisante pour permettre l'émergence des mots nouveaux, car leurs fréquences *a priori* faibles les pénalisent face aux «anciennes» formes dont les fréquences ont été estimées sur des textes de plusieurs centaines de millions de mots. La solution présentée ici est de développer des méthodes automatiques d'adaptation utilisant des textes temporellement proches des documents à transcrire [Auzanne *et al.*, 2000; Kemp and Waibel, 1998; Federico and Bertoldi, 2001; Allauzen and Gauvain, 2003]. Ces textes diffusés sur l'internet par des sites mis à jour quotidiennement sont utilisés pour modéliser le contenu lexical et linguistique de l'actualité. Ce chapitre est consacré à l'adaptation du vocabulaire, et le chapitre 8 à l'adaptation du modèle de langage.

Trois méthodes d'adaptation automatique du vocabulaire sont proposées. La première dite *ad hoc*, fait l'hypothèse de la disponibilité d'un vocabulaire de référence. C'est une heuristique qui utilise des seuils sur les fréquences de mots pour identifier les nouveaux mots à inclure dans le vocabulaire. Les résultats obtenus valident l'approche. Cependant les limites fonctionnelles de cet algorithme incitent à une formalisation théorique du problème. Ainsi le deuxième algorithme construit un vocabulaire de N mots à partir de l'interpolation linéaire de plusieurs distributions

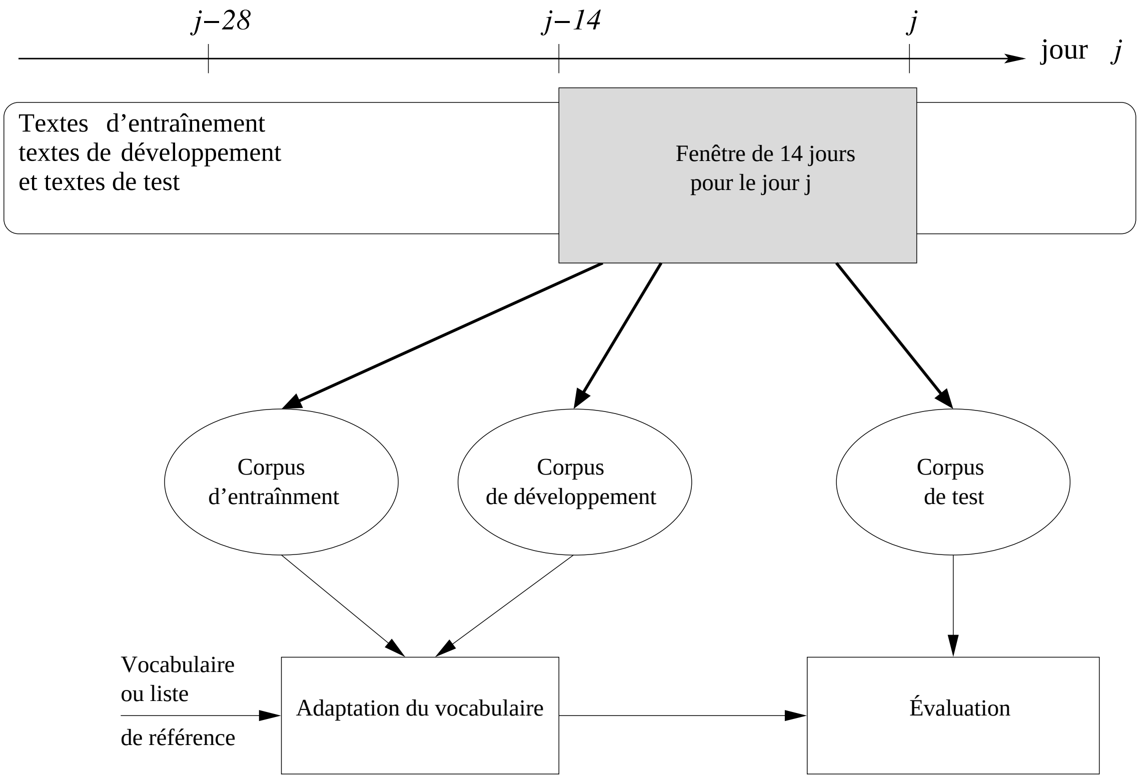


FIG. 6.1 – Cadre expérimental pour l'évaluation des méthode d'adaptation du vocabulaire. Les corpus sont constitués en appliquant une fenêtre de 14 jours sur le flux de textes extrait du Web. Par la suite, le corpus de test est noté $T_{14}(j)$, le corpus de développement $D_{14}(j)$, et le corpus d'apprentissage est $A_{14}(j)$.

sur des corpus distincts en optimisant la perplexité d'un corpus de développement. Et le troisième optimise la couverture lexicale sur un corpus de développement selon un critère propre au formalisme. Le même cadre expérimental représenté à la figure 6.1 a été mis en place pour évaluer les trois méthodes. Pour chaque jour du mois de janvier 2002, trois corpus sont constitués à partir des textes du Web : un corpus d'entraînement, un de développement et un de test. Les vocabulaires sont évalués en mesurant leur couverture lexicale sur les corpus de test. L'adaptation du vocabulaire est aussi évaluée en terme de taille du vocabulaire pour une couverture lexicale souhaitée.

6.1 Corpus extrait du Web

De nombreux sites Internet permettent de suivre l'actualité quasiment en temps réel. Certains contiennent une forme résumée de quotidiens nationaux, et d'autres diffusent des dépêches d'agences de presse. À partir d'un ensemble de sites, il est donc possible de collecter quotidiennement des textes reflétant le contenu lexical et linguistique des documents d'actualité.

6.1.1 Normalisation des textes

Les pages au format HTML collectées chaque jour sur les sites choisis sont "nettoyées" automatiquement par une procédure spécifique à chaque source. Ces documents sont tout d'abord convertis dans un format unique pour ne garder que la date, les titres et le texte des articles. Le codage des caractères est uniformisé et les textes sont ensuite normalisés. La normalisation est une étape majeure dans la préparation du corpus d'entraînement qui a déjà été évoquée au paragraphe 4.4.2. L'objectif est ici différent : la normalisation doit permettre une forte discrimination afin d'identifier les mots nouveaux tout en restant robuste aux bruits et aux erreurs des sources de textes. La solution usuelle pour rester robuste est d'utiliser une liste de termes de référence comme par exemple le vocabulaire du système de reconnaissance [Barras *et al.*, 2002]. Cependant, se référer à une liste d'autorité ne permet pas l'émergence de mots nouveaux. La normalisation est donc organisée comme dans [Allauzen and Gauvain, 2002] en deux étapes et reprend en grande partie les étapes décrites au paragraphe 4.4.2.

La première étape indépendante du lexique débute par le traitement des unités fréquentes (la même unité se rencontre sous des formes typographiques diverses qu'il est nécessaire de normaliser), et la réécriture d'abréviations usuelles. La segmentation en mots prend en compte les séparateurs ambigus tels que l'apostrophe et le tiret hors mots composés, les ponctuations et le cas particulier des nombres ordinaux, dates, heures, et prix. Cette première étape s'achève par le filtrage des adresses postales, des adresses Internet et certaines incises. La seconde étape s'appuie sur le vocabulaire de référence pour déterminer la casse de la première lettre de chaque phrase et pour décider s'il est nécessaire de séparer les mots composés. La normalisation des antonomases (également désignées par "majuscules emphatiques") n'a pas été jugée nécessaire car elle implique une intervention manuelle importante : le fait de savoir si une majuscule en début de mot relève d'une antonomase, d'une erreur typographique, ou traduit un nom propre, nécessite une analyse syntaxique, voire sémantique [Adda and Adda-Decker, 1997]. Finalement, les nombres sont convertis en mots, et seules les phrases de plus de huit mots sont conservées afin de filtrer les énumérations et les phrases structurelles induites par le type de texte manipulé.

6.1.2 Constitution des corpus

Cinq sites ont été sélectionnés pour leur contenu et leur facilité de téléchargement. Le tableau 6.1 résume dans sa partie supérieure, le comportement de chaque source par la valeur moyenne, minimale et maximale du nombre de mots collectés par jour sur l'ensemble des mois de décembre 2001 et janvier 2002. Force est de constater que la quantité de données fournie par chaque site est très variable (par exemple certaines sources peuvent ne diffuser aucun document nouveau les dimanches ou jours fériés). La source T fournit chaque jour un résumé de l'actualité et constitue notre corpus d'évaluation. Sur les quatre autres sites (D, $A^{(1)}$, $A^{(2)}$, $A^{(3)}$), deux sont des sites de quotidiens nationaux d'actualité et deux diffusent en continu des dépêches d'agences de presse. Chacun offre *a priori* un traitement de la même actualité (mêmes faits et mêmes thèmes) avec un point de vue éditorial qui lui est propre. Pour que les corpus d'entraînement, de développement et de test couvrent la même actualité, ces corpus sont constitués à partir d'un ou plusieurs de ces sites : le corpus de test est construit à partir de la source T, le corpus de

Source	*Moyenne*	*Min*	*Max*
T	6k	0k	11k
D	23k	0k	38k
$A^{(2)}$	58k	1k	100k
$A^{(1)}$	24k	14k	32k
$A^{(3)}$	45k	21k	72k
T_{14}	76k	62k	88k
D_{14}	152k	100k	176k
A_{14}	1 359k	1 054k	1 550k
A_{28}	2 650k	2 546k	2 947k

TAB. 6.1 – Moyenne, minimum et maximum du nombre de mots par jour et par source observés pour les mois de décembre 2001 et janvier 2002.

développement à partir de D, et le corpus d'entraînement à partir des sources $A^{(1)}$ à $A^{(3)}$. Par la suite, $S(j)$ désigne l'ensemble des textes du jour j venant du site S.

Pour constituer des ensembles de textes de taille suffisante et homogène, il est nécessaire de considérer plusieurs jours consécutifs. Pour chaque jour j du mois de janvier 2002, $A_k(j)$ désigne le corpus d'adaptation contenant les textes des jours j-k à k : $A_k(j) = \bigcup[A(j-k) \cdots A(j)]$ avec $A(j) = \bigcup_i A^{(i)}(j)$. De même, le corpus de développement pour le jour j comprend les textes $D(j)$ des jours j-k à j : $D_k(j)$.

Nous avons choisi k de telle sorte que le corpus d'adaptation comprenne au moins 2,5 millions de mots. Pour le corpus de test, une image précise de l'actualité est recherchée, mais également une taille conséquente pour que les mesures ne soient pas trop bruitées. Un corpus d'au moins 60 000 mots constitue un bon compromis. En prenant en compte ces contraintes, k a été fixé à 28 jours pour les données d'adaptation et à 14 jours pour les données de test. Comme le corpus de développement est utilisé pour l'adaptation lexicale, la valeur de k est la même que pour le corpus de test afin que les deux corpus couvrent la même période. La partie inférieure du tableau 6.1 montre que $T_{14}(j)$, c'est-à-dire l'ensemble des textes de test du jour même et des 13 derniers jours, contient en moyenne 76 k mots, l'ensemble des textes de développement $D_{14}(j)$ représente en moyenne 152,2 k mots, et $A_{28}(j)$, l'ensemble des textes d'entraînement du jour même et des 27 derniers jours, contient en moyenne 2,6 M mots.

6.2 Vocabulaire de référence

La construction du vocabulaire de référence est décrite en section 5.2. Rappelons juste que les données qui ont servi à sa construction datent des années 1987 à 1999, qu'il contient 65 333 mots, et qu'il a un taux de MHV de 1,2% sur un corpus de développement de l'année 2000.

Le pourcentage de MHV pour le vocabulaire de référence sur $T_{14}(j)$ varie suivant les jours entre 1,9% et 2,3% avec une valeur moyenne de 2,15% (voir figure 6.2). Ces valeurs sont à comparer avec le taux de 1,2% obtenu lors de l'optimisation du vocabulaire. Considérons l'ensemble du corpus d'évaluation pour le mois de janvier, sur lequel nous avons effectué manuellement une typologie des MHV. Dans ce corpus, 45% des MHV correspondent à des entités nommées (noms de personnes, de lieux, d'organisation), 20% sont dus à des erreurs typographiques, 20% sont des noms communs (en grande majorité des formes fléchies de mots déjà dans le vocabulaire) et 10% sont imputables à la normalisation de la première lettre des mots (principalement des antonomases). De plus, environ un tiers des formes MHV (34%) n'apparaissent qu'un seul jour.

6.3 Adaptation *ad hoc*

Pour diminuer le taux de MHV, l'adaptation *ad hoc* ajoute des mots au vocabulaire de référence, et donc en enlève d'autres pour rester dans les limites des 65 535 mots. Ce faisant, il faut tenir compte du fait que les corpus d'adaptation sont issus de sites Internet plus "bruités" que des textes de journaux.

6.3.1 Heuristique pour la mise à jour du vocabulaire

Afin de filtrer les éventuelles erreurs provenant des sources ou de la normalisation, un seuil est appliqué au nombre d'occurrences de chaque forme. Que ce soit pour les journaux radio et télédiffusés, ou pour la presse papier, le contenu de l'actualité est un ensemble de thèmes dont les durées de vie peuvent être considérées comme aléatoires. Pour un jour j, les journaux traitent de thèmes ponctuels qui n'apparaissent que ce jour, mais aussi de thèmes récurrents présents dans l'actualité depuis plusieurs jours. En prenant $A_1(j)$ comme corpus d'adaptation (c'est-à-dire uniquement les textes du jour j), les thèmes ponctuels y sont bien représentés, mais la quantité de données est très réduite. Par contre, en prenant $A_{28}(j)$ (les textes sur les quatre semaines précédentes), les thèmes récurrents masquent les thèmes ponctuels.

Nous notons $E_n(A_k(j))$, l'ensemble des mots apparaissant plus de n fois dans $A_k(j)$ et n'appartenant pas au vocabulaire de référence. Selon la loi de Zipf (pour plus de détails voir le parapgraphe 4.1.1), dans un corpus suffisamment grand, la fréquence d'un mot est proportionnelle à l'inverse de son rang. Ceci implique que peu de mots ont des fréquences élevées et qu'inversement beaucoup de mots apparaissent peu de fois. Pour n faible, le nombre de mots apparaissant n fois est très important. Il n'est donc pas possible de déterminer *a priori* une quantité de mots nouveaux à ajouter, car cela nécessite, pour un seuil n, de ne retenir qu'une partie des formes pour atteindre le nombre de mots fixé. La sélection des mots serait alors en partie arbitraire. Le choix du seuil a donc ici un impact prépondérant sur la quantité de mots nouveaux. La valeur de n doit être choisie de manière à obtenir suffisamment de mots nouveaux tout en rejetant les erreurs typographiques. Ainsi, l'ensemble des mots candidats à l'entrée (*in*) dans le vocabulaire pour le jour j ($E_{in}(j)$) est obtenu de la façon suivante :

$$E_{in}(j) = E_5(A_{28}(j)) \cup E_1(A_1(j)) \tag{6.1}$$

c'est-à-dire en faisant l'union des MHV apparaissant plus d'une fois dans $A_1(j)$ et des MHV apparaissant plus de 5 fois dans $A_{28}(j)$. Le nombre de mots dans $E_{in}(j)$ est noté $N_{in}(j)$.

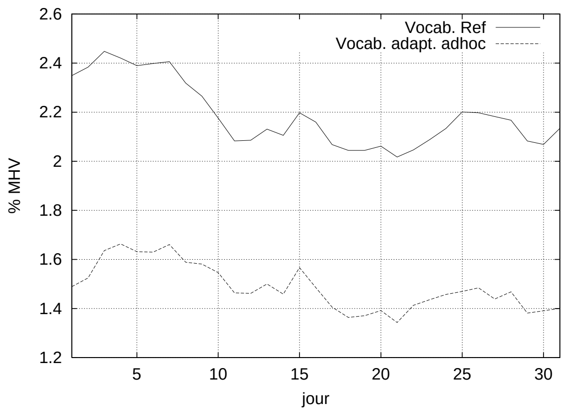


FIG. 6.2 – % de MHV pour chaque jour j sur $T_{14}(j)$ avec le vocabulaire de référence et le vocabulaire adapté.

Comme la taille du vocabulaire doit rester constante (65k mots), nous devons pour chaque candidat à l'entrée choisir un candidat en sortie. Pour cela, le vocabulaire de référence est trié dans l'ordre des probabilités unigrammes fournies par le modèle de langage de référence. Pour éviter de retirer un mot de français général, les 30 000 mots les plus fréquents ne peuvent être des candidats à la sortie du vocabulaire ainsi que tous les mots apparaissant dans $A_{28}(j)$. Les $N_{in}(j)$ mots ayant les probabilités unigrammes les plus faibles sont sélectionnés pour former la liste des candidats à la sortie du vocabulaire. Comme au chapitre précédent, si le nombre de mots à retirer est imposé, la présence éventuelle de mots équiprobables peut entraîner une sélection arbitraire. Cependant, le modèle de langage de référence résulte de l'interpolation de trois modèles comme décrit au chapitre 4. Ceci réduit le nombre d'ex aequo à des proportions négligeables lorsque l'on cherche à exclure plus de 1 500 mots. On obtient pour chaque jour j un vocabulaire de 65 333 mots nommé $V_{adhoc}(j)$ en remplaçant les mots candidats à la sortie par les mots candidats à l'entrée.

6.3.2 Expérimentation

La figure 6.2 permet de visualiser l'évolution du pourcentage de MHV sur $T_{14}(j)$ obtenu pour le vocabulaire de référence et le vocabulaire adapté $V_{adhoc}(j)$. L'adaptation du vocabulaire permet, par rapport aux résultats obtenus avec le vocabulaire de référence en terme de taux de MHV, un gain absolu qui varie entre 0,62% et 0,86% avec une moyenne de 0,72% soit une diminution relative moyenne de 32%. Le taux de MHV pour le vocabulaire adapté est très proche du taux de 1,2% qui avait été obtenu avec le vocabulaire de référence sur son corpus de développement. La proportion en nombre d'occurrences des mots retirés du vocabulaire et apparaissant dans $T_{14}(j)$, que nous appelons ici taux de réjection, ne dépasse pas 0,04%. L'algorithme d'actualisation du vocabulaire ajoute entre 1 486 et 2 114 mots par jour avec une moyenne de 1 728 mots.

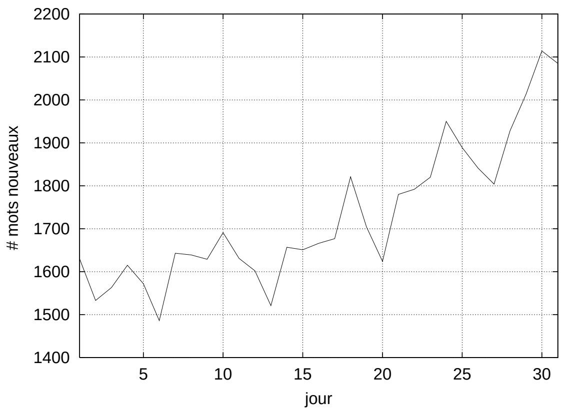


FIG. 6.3 – Nombre de mots nouveaux pour chaque jour

6.3.3 Analyse quantitative des mots ajoutés

Une des ambitions de l'algorithme est de sélectionner à la fois des mots nouveaux liés aux thèmes ponctuels et des mots nouveaux liés aux thèmes récurrents grâce à l'utilisation de deux listes. S'il est difficile d'apprécier qualitativement si les mots ajoutés sont "ponctuels" ou "récurrents", il est possible de quantifier l'apport de chaque liste : en moyenne $E_1(A_1(j))$ permet d'identifier 240 mots nouveaux dont 113 ne figurent pas dans les textes du mois $E_5(A_{28}(j))$. Ces 113 formes représentent 7,5% du nombre total de mots nouveaux ajoutés.

La figure 6.3 montre que le nombre de mots ajoutés a tendance à augmenter avec le temps. La courbe regardée à l'échelle de la semaine montre un minimum local pour chaque dimanche, les sources Internet de textes étant moins fournies le dimanche (les dimanches du mois de janvier 2002 sont les jours 6, 13, 20, 27).

L'ensemble des mots ajoutés au mois de janvier, comptent 6 400 formes distinctes. Seuls 369 mots (6%) sont présents sur l'ensemble du mois et 3 209 n'apparaissent qu'une fois.

Le tableau 6.2 comprend les 20 mots nouveaux ayant obtenu les probabilités unigrames les plus élevées sur le mois de janvier. Sur ces 20 mots, la moitié sont des noms propres (Laden, Enron, Duhalde, Rua, Al-Qaïda, Al-Qaida, Karzaï, Musharraf, Poutine, Tora). Le nom "Al-Qaïda" est présent sous deux orthographes différentes. Les autres formes précédemment incluses dans le vocabulaire de référence se distinguent ici par la majuscule qui relève de l'antonomase transformant ainsi ces noms communs en des formes comparables à des entités nommées. La majuscule fait référence à des contextes particuliers tels que : "Conseil constitutionnel", "Premier Ministre", et "Autorité palestinienne".

6.3.4 Pérennité de l'adaptation *ad hoc*

Afin de caractériser plus précisément la vitesse d'évolution lexicale de l'actualité et la pérennité de cette adaptation *ad hoc*, les taux de MHV sur $T_{14}(j)$ des trois vocabulaires suivants

1	Laden	-3,2998	11	Commission	-3,8459
2	Conseil	-3,4422	12	Banque	-3,8647
3	Premier	-3,4970	13	Etats	-3,8826
4	Autorité	-3,6254	14	Duhalde	-3,9198
5	Tora	-3,7346	15	Al-Qaida	-3,9328
6	Karzaï	-3,7445	16	Musharraf	-4,0205
7	Assemblée	-3,7610	17	Rua	-4,0360
8	Enron	-3,7635	18	Poutine	-4,1171
9	Al-Qaïda	-3,7716	19	Congrès	-4,1223
10	Cour	-3,7885	20	Centre	-4,1419

TAB. 6.2 – Les 20 premiers nouveaux mots de $V_{adhoc}(j)$ classés par leur probabilité unigramme pour le mois de janvier 2002. Les probabilités sont données sous la forme logarithmique en base 10.

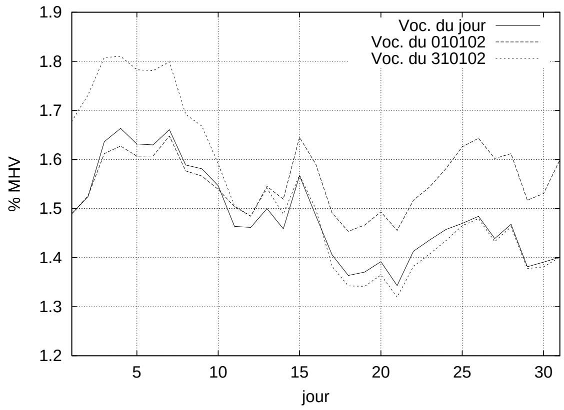


FIG. 6.4 – Comparaison du taux de MHV sur $T_{14}(j)$ obtenu avec 3 vocabulaires particuliers : $V_{adhoc}(j)$, $V_{adhoc}(01)$, $V_{adhoc}(31)$.

sont comparés sur la figure 6.4 : le vocabulaire adapté quotidiennement $V_{adhoc}(j)$, le vocabulaire adapté pour le premier janvier $V_{adhoc}(01)$ et le vocabulaire adapté pour le 31 janvier $V_{adhoc}(31)$. Du premier au 10 janvier, les taux de MHV avec $V_{adhoc}(j)$ et $V_{adhoc}(01)$ sont sensiblement identiques. Alors qu'avec $V_{adhoc}(31)$, il est supérieur de 0,2% au premier janvier (11% relatif). Cette différence décroît jusqu'à s'annuler aux environs du 10 janvier. À partir de cette date, la tendance s'inverse. Les performances de $V_{adhoc}(01)$ se dégradent en comparaison avec les autres vocabulaires jusqu'à atteindre une différence de 0,2% également au 31 janvier (12% en relatif).

Ces expérimentations ont été menées avec pour seul corpus d'entraînement $A_{28}(j)$, afin de comparer cet algorithme à celui présenté au chapitre suivant, avec les mêmes corpus d'entraînement et de test. Les mêmes expérimentations, mais incluant $D_{28}(j)$ dans le corpus d'entraînement, ont donné des résultats similaires.

6.3.5 Faiblesses de l'algorithme

L'algorithme décrit en section 6.3 est une heuristique qui hiérarchise *a priori* les listes de mots utilisées. Il requiert la définition préalable d'une liste de référence, puis d'une liste complémentaire permettant de sélectionner les mots nouveaux. L'extension de cet algorithme à plus de deux listes n'est guère envisageable car il faudrait être en mesure, soit d'établir un critère pour hiérarchiser ces listes, soit de pouvoir réunir ou fusionner plusieurs listes en une seule. De plus, l'adaptation *ad hoc* n'utilise pas de corpus de développement, il faut sélectionner manuellement la source ponctuelle $A_{28}(j)$ à utiliser. Les seuls paramètres sont les seuils de décision. Or ces paramètres sont fixés par des considérations de filtrage et sont indépendants de l'application visée. L'adaptation *ad hoc* se restreint à la mise à jour *a priori* d'un vocabulaire à l'aide d'une distribution de mots. Dans la suite, deux nouveaux algorithmes sont proposés. Le nombre de corpus d'entraînement (ou de listes) est libre et l'introduction d'information *a priori* se fait par l'utilisation d'un corpus de développement.

6.4 Adaptation par interpolation linéaire de modèle unigramme

Au paragraphe 5.3.4, un algorithme de construction automatique du vocabulaire est proposé, qui interpole un ensemble de modèles unigrammes estimés sur les données d'entraînement, afin de minimiser la perplexité du corpus de développement. Cet algorithme peut également être appliqué à l'adaptation du vocabulaire.

6.4.1 Algorithme d'adaptation du vocabulaire

Un modèle unigramme standard est adapté par une interpolation linéaire avec un modèle « local » propre au jour j. Si les probabilités attribuées au mot w, respectivement par les modèles standard, local et interpolé, sont notées $P_{std}(w)$, $P_{interpol}(w)$ et $P_{local}(w)$, alors l'interpolation linéaire s'écrit :

$$P_{interpol}(w) = \lambda P_{local}(w) + (1 - \lambda) P_{std}(w) \tag{6.2}$$

Le coefficient d'interpolation est calculé de manière à minimiser la perplexité du corpus de développement $D_{14}(j)$ via l'algorithme E.M [Dempster *et al.*, 1977]. Deux possibilités ont été expérimentées pour le modèle standard, soit le modèle unigramme de référence, soit le modèle unigramme obtenu au chapitre 5 à partir de la segmentation des corpus référencée « Années et type de source »[1]. Le modèle local est estimé sur le corpus d'entraînement $A_{28}(j)$. Le vocabulaire est construit à partir de cette distribution interpolée en choisissant les $N = 65\ 333$ mots les plus probables. La taille N du vocabulaire est la même que la taille du vocabulaire de référence ou du vocabulaire *ad hoc*.

[1]Cette segmentation sépare les textes selon le type de leurs sources (presse écrite, transcription d'extraits d'émissions, et transcription) et suivant les années. Plus de précisions peuvent être trouvées au 5.4.3.

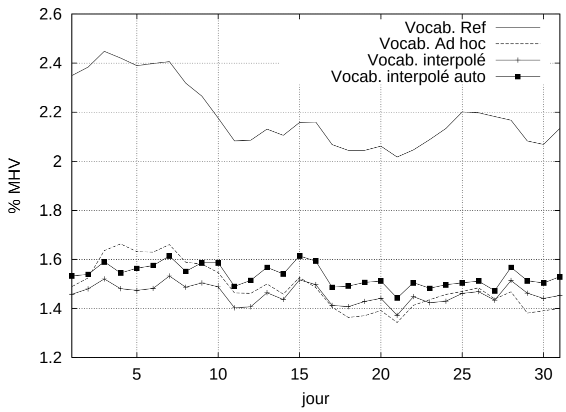


FIG. 6.5 – % de MHV pour chaque jour j sur $T_{14}(j)$ avec le vocabulaire de référence et le vocabulaire adapté selon la méthode *ad hoc* et les deux vocabulaires adaptés par interpolation linéaire de modèle unigramme. Le vocabulaire référencé « Vocab. interpolé » utilise comme modèle standard le modèle de référence alors que le vocabulaire référencé « Vocab. interpolé auto » utilise le modèle « Années et nature » généré automatiquement.

6.4.2 Expérimentation

La figure 6.5 permet de visualiser l'évolution du pourcentage de MHV sur $T_{14}(j)$ obtenu pour le vocabulaire de référence et les deux vocabulaires adaptés par interpolation linéaire de modèles unigrammes. Le taux de MHV obtenu avec le vocabulaire *ad hoc* est également rappelé. Par rapport au vocabulaire de référence, les gains sont tout à fait comparables à celui obtenu avec le vocabulaire *ad hoc* puisqu'ils sont respectivement de 33% et de 30% en moyenne (gains relatifs), suivant que le modèle standard utilisé soit le modèle de référence ou le modèle « Années et type de source ». L'utilisation du vocabulaire de référence donne en moyenne de meilleurs résultats. Le coefficient d'interpolation attribué au modèle standard est à peu près constant et égale à 0,4. Notons également que cet algorithme introduit en moyenne 16 645 mots nouveaux lorsque le modèle de référence est utilisé.

Cependant la méthode s'avère décevante puisque les différences de performances entre les trois vocabulaires adaptés ne sont pas probantes : le taux de MHV moyen sur le mois de janvier sont de 1,48 pour le vocabulaire *ad hoc*, 1,45 si le modèle de référence est utilisé, et 1,53 avec le modèle « Années et type de source ». Ces résultats sont d'autant plus décevants, que l'algorithme s'était montré efficace pour la construction automatique du vocabulaire (expériences du chapitre 5).

Une manière plus radicale d'appliquer cet algorithme consiste à reconstruire complètement le vocabulaire en utilisant l'ensemble des corpus d'entraînement décrits au paragraphe 5.1.1, en y ajoutant le corpus d'entraînement du jour $A_{28}(j)$ et en utilisant le corpus de développement $D_{14}(j)$. En dehors d'un temps de calcul long (interpolation d'une vingtaine de modèles), les expériences ont donné des résultats similaires.

6.5 Adaptation vectorielle du vocabulaire

Un troisième algorithme est proposé, qui à partir d'un ensemble de corpus d'entraînement, optimise la couverture lexicale sur un corpus de développement par interpolation linéaire de listes (mots, fréquences). D'un point de vue fonctionnel, cet algorithme est équivalent au précédent puisqu'en entrée, le nombre de corpus d'entraînement n'est pas limité, et que la sortie est une suite de mots ordonnés suivant leurs fréquences.

6.5.1 Description de l'algorithme vectoriel

Comme en recherche d'information [Salton and McGill, 1997], il est possible dans notre cas de représenter un document, un corpus ou plus généralement une distribution de mots par un vecteur dans l'espace des mots : un mot est un axe qui est gradué par une mesure quantifiant l'importance du mot dans cette distribution. La mesure choisie est la plus élémentaire : la fréquence du terme. Contrairement à l'usage en recherche d'information, la fréquence du terme n'est pas pondérée par l'inverse de la fréquence des documents contenant ce terme, car pour construire un vocabulaire, tous les termes doivent être pris en compte, et pas uniquement ceux qui sont porteurs d'information.

Ainsi les K corpus d'entraînement (X_1, X_2, ... , X_K) et le corpus de développement Y deviennent les vecteurs ($\vec{X_1}$, ... , $\vec{X_K}$) et $\vec{Y}$. Si les mots de $\vec{Y}$ sont triés par ordre décroissant de leurs occurrences et que les N mots les plus fréquents sont sélectionnés, on obtient bien évidemment le vocabulaire de N mots ayant la meilleure couverture lexicale sur $\vec{Y}$. L'objectif de l'adaptation vectorielle est d'estimer $\vec{Y}$ par combinaison linéaire des vecteurs d'entraînement.

Si la famille $(\vec{X_1}, ..., \vec{X_K})$ est libre, alors elle forme une base d'un sous-espace $\mathcal{S}$ de dimension K. Il y a alors deux possibilités, la première est que $\vec{Y}$ appartient à $\mathcal{S}$. La seconde (celle qui sera toujours rencontrée en pratique) est que $\vec{Y}$ n'appartient pas à $\mathcal{S}$. Dans ce cas, la meilleure approximation de $\vec{Y}$ est son projeté orthogonal $\vec{Y_p}$ dans $\mathcal{S}$.

Une condition suffisante pour vérifier que $(\vec{Y}, \vec{X_1}, ..., \vec{X_K})$ est libre est que chaque vecteur contienne un mot que les autres n'ont pas. Le but est donc de calculer les coordonnées de $\vec{Y_p}$, le projeté orthogonal de $\vec{Y}$ sur $\mathcal{S}$, dans la base $(\vec{X_1}, ..., \vec{X_K})$. La procédure suivante est utilisée :

1. L'indépendance linéaire de la famille $(\vec{X_1}, ..., \vec{X_K}, \vec{Y})$ est vérifiée à l'aide de la condition suffisante précédemment énoncée. Si un des vecteurs $\vec{X_i}$ n'est pas linéairement indépendant, $(\vec{X_1}, ..., \vec{X_K})$ n'est pas une base. L'algorithme s'arrête et il faut supprimer la ou les entrées inutiles. Si $\vec{Y}$ n'est pas linéairement indépendant[2], $\vec{Y} = \vec{Y_p}$.
2. Une base orthogonale $(\vec{x_1}, ..., \vec{x_K})$ de $\mathcal{S}$ est construite.
3. La projection de Y dans b se calcule de la façon suivante :

$$\vec{Y_p} = \sum_{i=1}^{K} \frac{\vec{Y}^t \vec{x_i}}{\| \vec{Y} \| \| \vec{x_i} \|} \vec{x_i} \ . \tag{6.3}$$

[2]En pratique, ce cas ne se produit jamais.

4. Un changement de base ($x \rightarrow X$) est effectué pour obtenir les coefficients $(\alpha_i)_{i=1}^{K}$ tels que :

$$\vec{Y_p} = \sum_{i=1}^{K} \alpha_i \vec{X_i} \ . \tag{6.4}$$

Le vocabulaire est finalement construit en sélectionnant les mots ayant les comptes interpolés les plus élevés dans $\vec{Y_p}$.

Aucune connaissance *a priori* sur les corpus d'entrée, leur taille ou sur leurs relations n'est nécessaire. Ceci permet d'utiliser en entrée des vecteurs de tailles très différentes sans pour autant que l'un soit occulté par l'autre. Cependant, le sous-espace de projection $\mathcal{S}$ est fixé par le choix des vecteurs d'entrées. L'algorithme dépend donc du choix des entrées par l'intermédiaire du sous-espace qu'elles engendrent.

6.5.2 Expérimentation

Les fréquences du corpus de développement $D_{14}(j)$ constituent le vecteur cible $\vec{Y}$. Le vecteur $\vec{X_1}$ est la liste des fréquences de mots obtenus à partir des journaux *Le Monde* et *Le Monde Diplomatique* de 1987 à 1999, soit 332 M de mots. Ce vecteur est fixe et modélise le français général. Afin d'éliminer les erreurs typographiques, les mots apparaissant dix fois ou moins ont été rejetés lors de la constitution de $\vec{X_1}$ qui contient 191 117 formes différentes. Le vecteur $\vec{X_2}$ est construit sur $A_{14}(j)$ qui contient en moyenne de 1,3 million de mots, et ne peut alors être considéré comme représentatif du contenu lexical du français, mais comme une image partielle de l'actualité du jour. Un mot doit apparaître plus d'une fois pour être dans le vecteur et ce afin de filtrer un minimum de bruit. $\vec{X_2}$ est composé en moyenne de 54 215 formes différentes. L'algorithme donne une liste de comptes interpolés $\vec{Y_p}$ dont les 65k premiers mots sont conservés pour former le vocabulaire du jour $V_{vect}(j)$.

Sur la figure 6.6, est tracée l'évolution du taux de MHV pour le mois de janvier sur $T_{14}(j)$ obtenu avec le vocabulaire de référence, $V_{adhoc}(j)$, et $V_{vect}(j)$. L'adaptation vectorielle donne les meilleurs résultats, avec un gain absolu par rapport au vocabulaire de référence allant de 0,71% à 0,97% avec une moyenne 0,82% (soit en relatif allant de 32 à 41% avec une moyenne de 37%). Le vecteur $\vec{X_2}$ est construit sur $A_{14}(j)$ qui regroupe les textes d'entraînement couvrant la même période que celle du corpus de développement. Les mêmes expérimentations ont été faites avec $\vec{X_2} = A_{21}(j)$ et $A_{28}(j)$ sans donner des résultats significativement différents.

6.5.3 Analyse quantitative des mots ajoutés

Le nombre de mots ajoutés par jour est à peu près constant autour de sa valeur moyenne de 8 020. Sur l'ensemble des vocabulaires $V_{vect}(j)$ du mois de janvier, les mots nouveaux représentent 19 343 formes distinctes. L'examen de la fréquence avec laquelle chaque nouveau mot est ajouté au vocabulaire montre qu'une fréquence domine : 3 252 formes sont présentes tous les jours. Puis en second, 2 980 formes sont ajoutées 14 fois dans le mois montrant ainsi l'impact de la taille de la fenêtre utilisée pour le vecteur d'entrée $\vec{X_1}$ qui est justement de 14 jours. Et enfin seulement 751 formes n'apparaissent qu'une seule fois.

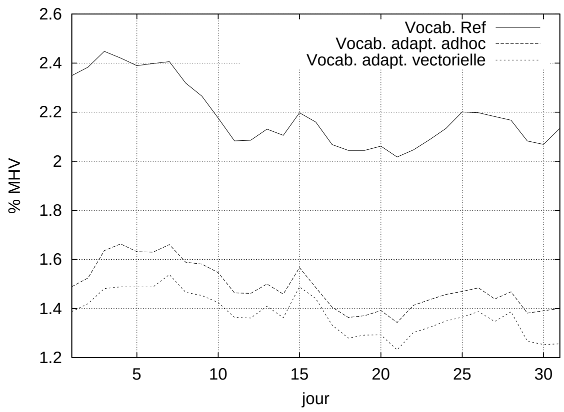


FIG. 6.6 – % de MHV pour chaque jour j sur $T_{14}(j)$ avec le vocabulaire de référence puis les 2 vocabulaires adaptés.

Le rang lexical d'un mot dans $V_{vect}(j)$ s'obtient en classant les mots du vecteur $\vec{Y_p}$ par ordre décroissant des fréquences interpolées. Le rang moyen d'un nouveau mot est 41 470, mais l'algorithme hisse localement des mots parmi les plus fréquents comme le montre le tableau 6.3. Ce tableau comprend les 20 mots nouveaux ayant obtenu les rangs lexicaux les plus élevés sur le mois de janvier. D'ailleurs, sur ces 20 mots, 12 sont des noms propres (Laden, Enron, Duhalde, Rua, Al-Qaïda, Saa, Karzaï, Musharraf, Zacarias, Perruche, Zinni, Poutine). Les autres formes précédemment incluses dans le vocabulaire de référence se distinguent ici par la majuscule qui relève de l'antonomase transformant ainsi ces noms communs en des formes comparables à des entités nommées. La majuscule fait référence à des contextes particuliers tels que : "Conseil constitutionnel", "Premier Ministre", et "Autorité palestinienne".

1	Conseil	169	11	Commission	681
2	Laden	194	12	Al-Qaïda	727
3	Premier	272	13	Saa	740
4	Enron	382	14	Karzaï	758
5	Assemblée	458	15	Musharraf	799
6	Autorité	509	16	Zacarias	968
7	Duhalde	527	17	Etats	969
8	Banque	547	18	Perruche	1 019
9	Cour	569	19	Zinni	1 033
10	Rua	633	20	Poutine	1 105

TAB. 6.3 – Les 20 premiers nouveaux mots classés par leur rang lexical dans $V_{vect}(j)$ pour le mois de janvier 2002

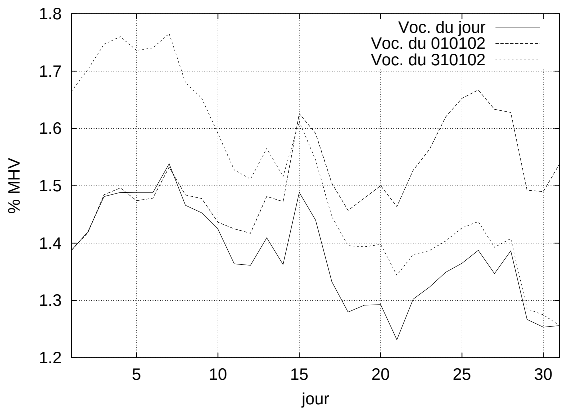


FIG. 6.7 – Comparaison du taux de MHV sur $T_{14}(j)$ obtenu avec 3 vocabulaires particuliers : $V_{vect}(j)$, $V_{vect}(01)$ et $V_{vect}(31)$.

6.5.4 Pérennité de l'adaptation vectorielle

Comme pour l'algorithme *ad hoc* (voir section 6.3.2 et figure 6.4), la figure 6.7 caractérise la pérennité de l'adaptation. Les taux de MHV sur $T_{14}(j)$ y sont représentés pour 3 vocabulaires : le vocabulaire adapté quotidiennement $V_{vect}(j)$, le vocabulaire adapté pour le premier janvier $V_{vect}(01)$, et le vocabulaire adapté pour le 31 janvier $V_{vect}(31)$. Les mêmes tendances que pour l'adaptation *ad hoc* sont observées. Au premier janvier, la différence relative entre $V_{vect}(31)$ et $V_{vect}(01)$ est de 17% en faveur de $V_{vect}(01)$ alors qu'au 31 janvier la différence est toujours de 18% mais cette fois-ci en faveur de $V_{vect}(31)$.

6.6 Comparaison des algorithmes *ad hoc* et vectoriel

La figure 6.6 montre que le vocabulaire $V_{vect}(j)$ donne les meilleurs résultats sur les données de test. Cependant entre les deux vocabulaires adaptés, le gain est faible sur l'ensemble du mois de janvier : 0,10% en absolu et 7% en relatif. De par sa construction, $V_{vect}(j)$ inclut beaucoup plus de nouveaux mots (8 000 contre 1 700 pour le vocabulaire *ad hoc*) et la sélection en entrée de l'algorithme est moins contraignante. Le risque d'inclure du bruit est donc plus élevé. Le taux de réjection de 0,12% (défini au 6.3.2) pour la méthode vectorielle est trois fois plus élevé que pour la méthode *ad hoc*. La comparaison des figures 6.4 et 6.7 montre qu'avec l'adaptation vectorielle, la différence relative entre le taux de MHV obtenu pour le vocabulaire adapté au premier janvier et celui obtenu pour le vocabulaire adapté au 31 janvier est de 18%, alors qu'elle est de 12% avec l'adaptation *ad hoc*. L'adaptation vectorielle est donc plus précise localement mais moins pérenne que l'adaptation *ad hoc*.

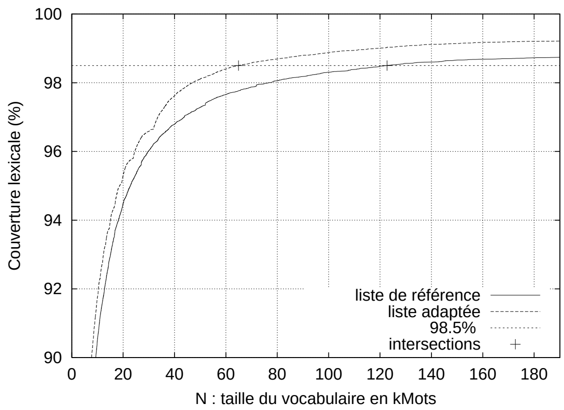


FIG. 6.8 – Couverture lexicale sur $T_{14}(15)$ en fonction de la taille du vocabulaire pour les listes $\vec{Y_p}$ et $\vec{X_1}$ (std)

6.7 Dimension du vocabulaire

La sélection des mots pour un système de reconnaissance de la parole se fait généralement en prenant les formes les plus fréquentes de plusieurs corpus d'apprentissage. L'objectif est de maximiser la couverture lexicale sur un corpus de développement pour une taille du vocabulaire N donnée. Dans la continuité des expériences du chapitre précédent, les deux vecteurs $\vec{Y_p}$ et $\vec{X_1}$ sont comparés par leur taux de couverture lexicale sur $T_{14}(j)$. Le vecteur $\vec{X_1}$ est la liste principale ayant servi à la construction du vocabulaire de référence, et $\vec{Y_p}$ est celle obtenue par l'algorithme d'adaptation vectorielle (décrit en section 6.5). La figure 6.8 représente la couverture lexicale en fonction de N pour le 15 janvier. La courbe de $\vec{Y_p}$ (ligne discontinue) est toujours au-dessus de celle de $\vec{X_1}$ (ligne continue) : $\vec{Y_p}$ engendre le vocabulaire ayant la meilleure couverture et ce pour tout N.

Plutôt que de travailler à N constant, on peut utiliser la couverture lexicale pour dimensionner le vocabulaire. Pour comparer la liste de référence et la liste adaptée, nous avons choisi une couverture lexicale supérieure ou égale à 98,5% (soit un taux de MHV de 1,5%). Sur la figure 6.8, cela correspond à la droite horizontale en pointillés. Les intersections de cette droite, avec les courbes représentant la couverture lexicale en fonction de la taille du vocabulaire obtenue avec $\vec{Y_p}$ et $\vec{X_1}$ pour le 15 janvier 2002, donnent les deux points de fonctionnement : pour une couverture de 98,5%, $\vec{Y_p}$ nécessite un vocabulaire de 64 908 mots et $\vec{X_1}$ un vocabulaire de 127 737. L'adaptation vectorielle permet donc de diviser par presque deux la taille du vocabulaire.

La figure 6.9 montre ce dimensionnement du vocabulaire pour l'ensemble du mois de janvier. L'adaptation permet une réduction de la taille du vocabulaire de 47 à 60% avec une moyenne de 50%. La droite horizontale en pointillés correspond à un vocabulaire limité à 65 535 mots. La couverture de la liste de référence est mauvaise au début du mois de janvier puisqu'il faut un vocabulaire de plus de 120 k mots pour les 10 premiers jours du mois. Par contre, la taille du vocabulaire adapté est stabilisée tout au long du mois aux environs des 65 535 mots.

Si la couverture lexicale exigée est de 99%, l'adaptation lexicale devient obligatoire. Car avec

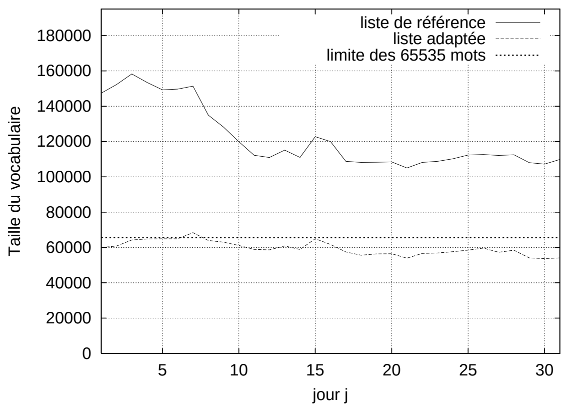


FIG. 6.9 – Évolution de la taille du vocabulaire nécessaire pour obtenir une couverture lexicale de 98,5% sur $T_{14}(j)$ avec les $\vec{Y_p}$ et $\vec{X_1}$

la liste $\vec{X_1}$, même en incluant les 191 117 formes, la couverture lexicale varie entre 98,6% et 98,9% avec une moyenne 98,75%. Par contre, la liste $\vec{Y_p}$ permet d'atteindre entre 99,4% et 99,1% avec un nombre de formes différentes de 193 060 mots en moyenne.

Conclusion

Afin de pouvoir suivre en temps réel les évolutions linguistiques et lexicales de l'actualité, des textes sont quotidiennement collectés à partir de sites internets. À partir de ces données, un cadre d'évaluation est mis en place comprenant un corpus d'entraînement, un autre de développement puis un corpus de test propre à chaque jour du mois de janvier 2002. Trois méthodes d'adaptation automatique du vocabulaire sont proposées et évaluées quotidiennement en terme de couverture lexicale. La première méthode *ad hoc* basée sur une heuristique valide l'approche et réduit le nombre d'occurrences de mots hors vocabulaire de 32 %. La deuxième s'inscrit dans le cadre théorique utilisé pour la construction automatique du vocabulaire (présenté au chapitre 5). Cependant, les résultats sont au mieux identiques à ceux obtenus avec la méthode *ad hoc*. Donc une troisième méthode qui utilise un formalisme vectoriel est développée et réduit le nombre d'occurrences de mots hors vocabulaire de 37 %. L'étude des performances de différents vocabulaires pour chaque jour du mois de janvier montre que les mots de l'actualité changent suffisamment rapidement pour justifier une adaptation quotidienne. De plus, le vocabulaire construit avec l'algorithme vectoriel est localement le plus performant mais est dépendant du corpus de développement, alors que le vocabulaire construit avec l'algorithme *ad hoc* conduit à une adaptation moins performante mais plus pérenne. Ces deux méthodes peuvent correspondre à deux usages donnés : la méthode *ad hoc* permet une sélection en aveugle, alors que la méthode vectorielle fait usage d'un corpus de développement. Ainsi lorsqu'un corpus de développement adéquat est disponible, la méthode vectorielle est la plus efficace.

Plutôt que de travailler avec un vocabulaire de taille fixe, la couverture lexicale peut être fixée pour dimensionner le vocabulaire. Les expériences montrent que l'adaptation vectorielle du vocabulaire permet en moyenne de diviser la taille du vocabulaire par deux pour une couverture lexicale de 98,5 %. Le vocabulaire comprend en moyenne 64,9 k mots contre 127,7 k mots sans adaptation. L'intégration de l'adaptation du vocabulaire dans un système de reconnaissance de la parole, ne peut se faire sans l'adaptation des modèles de langage qui fait l'objet du chapitre 8.

Chapitre 7

Construction d'un modèle de langage à vocabulaire ouvert

L'inconvénient d'un vocabulaire figé dépasse le cadre des émissions d'actualités. Les archives audiovisuelles contiennent des émissions de formes et de contenus très variés. Il est impossible de toujours disposer de données permettant de modéliser le contenu lexical et linguistique d'une émission. Il est donc utile de pouvoir introduire des mots nouveaux dans le système de reconnaissance, sans pour autant reprendre les étapes de construction du vocabulaire et du modèle de langage (cf. chapitres 4 et 5). Le modèle de langage et le système de reconnaissance automatique de la parole utilisent alors un vocabulaire ouvert.

Il est possible de prévoir une entrée supplémentaire dans le dictionnaire de prononciation, souvent appelé *filler model* [Bazzi, 2002]. Le système de reconnaissance automatique de la parole est alors capable de localiser une suite de phonèmes correspondant à un mot hors vocabulaire, et d'en donner une transcription phonétique. Les deux principaux inconvénients de cette approche est qu'elle nécessite une étape supplémentaire de conversion phonème-graphème, et que le *filler model* a tendance à générer des erreurs en absorbant des segments de parole qui ne concernent pas des mots hors vocabulaire. Le problème posé dans ce chapitre est différent. Il s'agit d'introduire dans le modèle de langage des mots hors vocabulaire qui sont connus à l'avance, c'est à dire que leurs graphies et leurs transcriptions phonétiques sont disponibles.

La méthode utilisée, dite du repli lexical est d'abord décrite. Puis, deux applications sont explorées. La première est appelée interactive car les mots nouveaux sont sélectionnés manuellement et introduits dans le vocabulaire (manière à supprimer la présence de mots hors vocabulaire). Cette expérimentation en oracle permet d'évaluer les capacités de la méthode à transcrire les mots ajoutés. La seconde est automatique, puisqu'elle utilise la méthode *ad hoc* (décrite au chapitre 6) pour sélectionner automatiquement les mots à ajouter. Les résultats de cette expérimentation dépendent de l'efficacité de la méthode *ad hoc* et des corpus d'entraînement qu'elle utilise. La première expérimentation correspond donc à une application où certains mots connus à l'avance n'appartiennent pas au vocabulaire. C'est le cas, lors de la transcription d'une émission pour laquelle, par exemple, les noms des invités sont préalablement connus. Par contre la seconde expérimentation peut être interprétée comme une nouvelle méthode d'adaptation du vocabulaire ne nécessitant pas une adaptation du modèle de langage.

7.1 Repli lexical

Pour qu'un mot existe du point de vue d'un système de reconnaissance automatique de la parole, il est indispensable qu'il fasse partie du vocabulaire, que sa transcription phonétique soit disponible, et que le modèle de langage soit apte à lui attribuer une probabilité non nulle. L'introduction d'un mot nouveau doit donc reprendre ces trois étapes. Inclure un mot nouveau dans le vocabulaire est une étape simple, et sa transcription phonétique peut être obtenue automatiquement. Par contre, les probabilités du modèle de langage sont estimées lors de son entraînement. Donc, pour inclure un mot nouveau sans ré-entraîner le modèle de langage, une forme est prévue lors de l'entraînement qui remplace un ou plusieurs mots nouveaux qui ne sont pas encore connus. Puis, lors du décodage, un mot nouveau peut être introduit comme une variante de prononciation de cette forme. Ainsi, l'accès à la probabilité du mot nouveau se fait par repli lexical de cette forme sur le mot. Cette forme peut être assimilée à une classe de mots, mais le terme « forme de repli » semble plus approprié.

La méthode du repli lexical se décompose donc en deux phases : l'entraînement du modèle et l'introduction de mots nouveaux via le lexique. Mais avant cela, une étude préliminaire est nécessaire afin de choisir ces formes de repli.

7.1.1 Choix des formes de repli

Le plus simple est de choisir une seule forme de repli. Cette forme existe déjà, c'est <UNK> qui modélise les mots hors vocabulaire (voir le paragraphe 4.4.3). Mais ce choix ne tient pas compte du type et du rôle des mots hors vocabulaire.

De plus la forme <UNK> a une probabilité surestimée, puisqu'elle regroupe les occurrences de tous les mots hors vocabulaire sur la même forme :

$$P(< UNK >) = \sum_{\text{mots hors vocabulaire}} P(\text{mot hors vocabulaire}).$$

L'étiquetage en catégories morpho-syntaxiques (*Part of Speech*) est utilisé pour différencier le rôle des mots. Un corpus d'observation est constitué à partir des corpus d'entraînement du modèle de langage qui sont décrits au paragraphe 4.4.1. Il contient six années du journal *Le monde* soit 103 360 165 mots. Le taux de mots hors vocabulaire calculé avec le vocabulaire de référence est de 1,54%, ce qui représente 1 591 282 occurrences. Ce corpus est étiqueté en catégories morpho-syntaxiques à l'aide du tagger de Brill[1] [Brill, 1994] adapté au français par l'INALF[2]. Lors de la campagne d'évaluation GRACE, ce tagger est celui qui a obtenue la meilleure précision avec 98%. Le tableau 7.1 rassemble les étiquettes les plus fréquentes concernant les mots

[1] Disponible à http ://www.cs.jhu.edu/~brill
[2] Disponible à http ://jupiter.inalf.cnrs.fr/WinBrill

Étiquette morpho-syntaxique	Fréquences	% des MHV
Nom propre	854043	53,7
Nom commun	266062	16,7
Verbe conjugué	157980	9,9
Adjectif	133363	8,4
Participe passé adjectif	91599	5,8
Verbe infinitif	34365	2,2
Adverbe	13486	0,8
Participe passé	9212	0,6
Nom commun étranger	4028	0,2
Participe présent	2167	0,1
Cardinal	1886	0,1

TAB. 7.1 – Fréquences sur le corpus d'observation des catégories morpho-syntaxiques pour les mots hors du vocabulaire de référence. La dernière colonne indique pour chaque catégorie morpho-syntaxique, le pourcentage des mots hors vocabulaire y appartenant.

hors vocabulaire[3]. Les noms propres représentent plus de la moitié des occurrences de mots hors vocabulaire, et la seconde catégorie des noms communs représente moins de 20%. Les catégories écartées concernent principalement des scories de la normalisation (par exemple issues du traitement des ponctuations, des sigles ou des incises), des abréviations, des néologismes ou des erreurs d'orthographes. L'ensemble de ces catégories représente moins de 0,2% et la contribution de la plus fréquente d'entre elles est de 0,05%.

L'ensemble des formes de repli est constitué à partir des étiquettes morpho-syntaxiques du présentées dans le tableau 7.1 auxquelles s'ajoutent la distinction de nombre. Pour les noms propres, toutes les étiquettes sont rassemblées en une seule classe, sans distinction de genre. Une classification supplémentaire est effectuée à l'aide de listes de noms géographiques et de prénoms collectés sur Internet. Les catégories les moins fréquentes n'ont pas été utilisées. Ainsi, un ensemble de 20 formes de repli est obtenu.

7.1.2 Construction du modèle de langage

Le modèle de langage est construit exactement comme le modèle de référence sauf en ce qui concerne le vocabulaire et la préparation des corpus. Le vocabulaire est construit en ajoutant au vocabulaire de référence (décrit à la section 5.2) les 20 formes de repli. La préparation des corpus d'entraînement permet de mettre en place la projection des textes sur le nouveau vocabulaire. Chaque mot du corpus d'entraînement est projeté sur une entrée du vocabulaire. Si le mot appartient au vocabulaire, il est projeté sur lui même. S'il n'appartient pas au vocabulaire alors il est projeté sur sa forme de repli, si celle-ci appartient au vocabulaire. Sinon, le mot est projeté sur la forme <UNK> qui regroupe les mots inconnus. Ceci est illustré par le tableau 7.2. En pratique, projeter est équivalent à remplacer.

[3] « mot hors vocabulaire » désigne toujours un mot hors du vocabulaire de référence. Il n'est jamais fait mention en ces termes du vocabulaire du tagger, qui est constitué des entrées du trésor de la langue française (le TLF est consultable en ligne sur le site http ://atilf.inalf.fr).

Kandinski et Brahem	sympathisaient	à Paris ou à Joyeuse
Kandinski et <SBP>	<VCJ :pl>	à Paris ou à <GEO>

TAB. 7.2 – Exemple de projection d'un texte d'entraînement sur le vocabulaire. Cet exemple suppose que les mots « Brahem », « sympathisaient » et « Joyeuse » sont des mots hors vocabulaire. $< SBP >$ est la forme de repli qui correspond aux noms propres quelconques, $< GEO >$ aux noms géographiques et $< VCJ : pl >$ aux verbes conjugués au pluriel

Le modèle de langage de référence quadrigramme contient environs 14 millions de bigrammes, 13 millions de trigrammes et 10 millions de quadrigrammes. Avec l'ajout des 20 formes de repli au vocabulaire, le nouveau modèle contient plus de 15 millions de bigrammes, 17 millions de trigrammes et 15 millions de quadrigrammes. Ainsi la taille du modèle a augmenté de 11,5%.

7.1.3 Introduction de mots nouveaux et phonétisation

Les mots nouveaux peuvent être introduits via le lexique de prononciation, sous forme de variantes de prononciation de la forme de repli correspondante. Chaque entrée lexicale est décrite à l'aide d'une séquence d'éléments choisis parmi 33 phonèmes, auxquels s'ajoutent 3 symboles spécifiques pour représenter les hésitations, les respirations et les silences. Pour un ensemble de 100 000 mots, les prononciations obtenues à partir d'un système de phonétisation automatique, ont été vérifiées et enrichies manuellement. Le système de phonétisation utilise des règles de conversion graphèmes-phonèmes [Prouts, 1980], auxquelles s'ajoutent des règles d'exception et des traitements particuliers pour les acronymes [de Mareüil, 1996]. Ce système a été classé parmi les premiers lors des évaluations relatées dans l'article [Yvon *et al.*, 1998].

La phonétisation des mots nouveaux est effectuée de la manière suivante : si la phonétisation est déjà disponible, elle est utilisée, sinon la prononciation est obtenue uniquement à partir de règles graphème-phonèmes. Le processus est donc automatique.

La probabilité associée à un mot introduit par la méthode du repli lexical s'écrit de la manière suivante :

$$P(w) = P(F(w))P(w|F(w)), \tag{7.1}$$

si $F(w)$ désigne la forme de repli associée au mot w. Le repli lexical se place dans un cadre d'application, où il n'existe pas de données d'adaptation concernant les mots à ajouter. L'estimation de la probabilité d'appartenance d'un mot à sa forme de repli $P(w|F(w))$ est alors problématique. Par la suite, cette distribution des mots au sein d'une forme de repli est donc uniforme.

7.2 Expérimentation

Deux cadres d'expérimentation sont envisagés selon le mode de sélection des mots introduits dans le vocabulaire : le premier est interactif, et le second automatique. L'expérimentation interactive est un oracle où les mots sont introduits manuellement, alors que dans l'expérimentation automatique, les mots introduits sont sélectionnés automatiquement grâce à la méthode

Etiquette morpho-syntaxique	%
Nom propre	58,3
Verbe conjugué	15,8
Nom commun	11,2
Participe passé adjectif	5,8
Adjectif	3,3
Verbe à l'infinitif	2,8
Participe passé	0,8
Adverbe	0,5
Participe présent	0,3

TAB. 7.3 – Proportion des étiquettes morpho-syntaxiques pour les 620 mots hors vocabulaire observés sur le corpus de test

ad hoc présentée à la section 6.3. Le corpus de test est décrit, puis les approches interactive, et automatique sont présentées.

7.2.1 Corpus de test

Le corpus de test est constitué de journaux télévisés diffusés par la chaîne TF1, recueillis tous les jours du mois de janvier 2002 au format RealVideo, à partir du site internet de la chaîne. Le premier quart d'heure de huit journaux répartis sur le mois a été transcrit manuellement avec le logiciel d'aide à l'annotation de signaux de parole *Transcriber* [Barras *et al.*, 2001]. De plus trois journaux télévisés, répartis sur le mois, ont été intégralement transcrits (le 3, 15 et 30 janvier). Le premier quart d'heure d'un journal télévisé représente en moyenne 2 850 mots et un journal complet 7 930 mots. La durée total du corpus de test est de 3h30.

Le vocabulaire de référence a un taux de mots hors vocabulaire de 1,21% sur ce corpus de test. Le corpus est étiqueté en catégorie morpho-syntaxique de la même manière qu'au paragraphe 7.1.1. La répartition des étiquettes pour les mots hors vocabulaire est donnée dans le tableau 7.3. Cette répartition est similaire à celle observée sur le corpus d'observation (tableau 7.1). Hormis quelques différences, l'ensemble des formes de repli reste valide puisqu'il est représentatif des catégories morpho-syntaxiques qui concernent les mots hors vocabulaire.

7.2.2 Expérimentation interactive

Il est parfois possible de connaître *a priori* quelques mots dont l'occurrence dans les données à transcrire est certaine. Cela peut être le cas pour les noms d'invités ou de personnes importantes dans l'actualité. Mais ces mots peuvent être hors vocabulaire. L'expérimentation interactive tente de répondre au besoin d'un système de reconnaissance dont le vocabulaire peut être modifié manuellement en fonction du corpus à transcrire en y introduisant des mots qui *a priori* seront prononcés.

L'objectif est donc d'évaluer l'aptitude de la méthode du repli lexical à permettre la reconnaissance des mots introduits, et non le mode de sélection des mots à ajouter aux vocabulaire.

			WER		
Jour	*#mots*	*%MHV*	*Réf*	*Repli*	*Ratio*
03	7828	1,35	28,0	24,3	2,7
04	2985	1,14	27,8	25,3	2,2
14	2867	1,36	20,4	18,4	1,5
15	8145	1,19	23,3	19,1	3,5
16	2813	1,28	21,9	20,5	1,1
23	2825	1,24	25,4	23,2	1,8
24	2846	1,09	27,9	24,7	2,9
28	2861	1,47	27,8	26,0	1,2
29	2835	1,02	26,8	24,8	2,0
30	7827	1,16	21,4	20,0	0,9
31	2754	0,80	20,1	19,2	1,1
total	46586	1,21	24,5	21,9	2,1

TAB. 7.4 – Les résultats de reconnaissance (WER) ont été obtenus avec d'une part le vocabulaire et le modèle de référence (Réf), et d'autre part, le vocabulaire et le modèle de langage utilisant le repli lexical (Repli). Le taux de MHV obtenu avec le vocabulaire de référence (%MHV) ainsi que le nombre de mots des émissions (#mots) sont également précisés. La dernière colonne (Ratio) est le rapport entre le gain absolu sur le WER et le taux de MHV.

Pour cela, le choix est fait dans cette expérimentation d'ajouter via le repli lexical tous les mots hors vocabulaire observés sur le corpus de test. Pour chacune des 11 émissions du corpus de test, la transcription de référence a été étiquetée par le même marqueur morpho-syntaxique que celui utilisé pour traiter les corpus d'entraînement (voir paragraphe 7.1.1). Les mots hors vocabulaire ont été identifiés ainsi que leur forme de repli associées, et ils sont introduits dans le vocabulaire comme cela est décrit au paragraphe 7.1.3.

Les résultats de reconnaissance sont rassemblés dans le tableau 7.4. Ce tableau montre que l'ajout manuel des mots hors vocabulaire par la méthode du repli lexical permet une amélioration relative du taux de d'erreur de 10,6%. Le gain absolu est en moyenne de 2,6% alors que le taux de mots hors vocabulaire moyen avant ajout est de 1,21%. Le rapport entre le gain sur le taux d'erreur et le taux de MHV est de 2,1. Ce résultat est en accord avec l'assertion : un mot hors vocabulaire entraîne en moyenne de 1,5 à 2 erreurs de reconnaissance.

Les résultats présentés précédemment n'impliquent pas que tous les mots nouveaux introduits dans le vocabulaire soient reconnus. Le tableau 7.6 rassemble pour chaque forme de repli, le nombre d'occurrences et le taux de reconnaissance observés sur le corpus de test. Le taux de reconnaissance moyen est de 78,7%. Une étude des erreurs sur les formes de repli montre qu'il est nécessaire de distinguer les noms propres, géographiques et les prénoms. Pour le reste, la phonétisation ne pose globalement pas de problème, et les sources d'erreurs sont « classiques » : segments écartés par le partionner, conditions acoustiques difficiles, parole superposée, et surtout des fautes d'accord imputables au modèle de langage. Pour les noms propres, les noms de lieu et les prénoms, les raisons précédentes restent valables, mais des problèmes de phonétisation automatique interviennent. Le système de phonétisation détecte la plupart du temps l'origine étrangère des noms propres comme « Limonov ». Lorsque l'origine n'est pas détectée, le mot est transcrit phonétiquement selon les règles françaises comme par exemple : « Vanbergue » qui est transcrit par /vAbErg/, ou « Prso » qui transcrit par PRSO. Sur l'ensemble des noms propres, noms de lieu et prénoms mal reconnus par le système, la phonétisation est responsable d'un tiers des erreurs.

Dans les autres cas, caractériser les erreurs est plus difficile, comme le montre l'exemple du tableau 7.5. Dans cette exemple typique, le nom propre « Izzo » est un mot nouveau qui est correctement phonétisé. Cependant la différence de perplexité entre l'hypothèse et la référence est de 8% en faveur de l'hypothèse (la log-vraisemblance est de -57 la référence et de pour -51 pour l'hypothèse) montrant ainsi que le modèle de langage contribue à l'erreur.

REF :	avec le personnage	*d' Izzo*	parce que j' ai beaucoup aimé	*ce qu' a*	*écrit Izzo*
HYP :	avec le personnage	*du jour*	parce que j' ai beaucoup aimé	*cette*	*crise ou*

TAB. 7.5 – Exemple d'erreur de reconnaissance sur un extrait un mot nouveau « Izzo ».

7.2.3 Expérimentation automatique

Au chapitre 6, différentes méthodes d'adaptations du vocabulaire ont été présentées. La méthode *ad hoc*, fait l'hypothèse de la disponibilité d'un vocabulaire de référence. C'est une heuristique qui utilise des seuils sur les fréquences de mots pour identifier les nouveaux mots à inclure dans le vocabulaire. La méthode vectorielle optimise la couverture lexicale sur un corpus de développement selon un critère vectoriel. Ces deux méthodes sont d'abord évaluées en terme de couverture lexicale sur le corpus de test : le vocabulaire adapté avec la méthode *ad hoc* obtient un taux de MHV de 1,04% au lieu de 1,21% (soit un gain relatif de 14%), tandis que le vocabulaire adapté suivant la méthode vectorielle obtient le même taux que le vocabulaire de référence. Les résultats différent de ceux du chapitre 6, puisque c'est la méthode *ad hoc* qui donne les meilleurs résultats. Cette différence dans les résultats est étudiée au chapitre 8. Pour cette expérimentation, la méthode *ad hoc* décrite à la section 6.3 est donc retenue.

Le gain en couverture lexicale est de fait nettement inférieur à celui obtenu dans le cadre de l'expérimentation interactive. Le gain sur le taux d'erreur devrait être au maximum de 0,5%, si on se réfère aux résultats précédents montrant un rapport moyen de 2,1 entre le gain en couverture lexicale et le gain sur le taux d'erreur. Cependant les expériences de reconnaissance montrent une dégradation des résultats avec une augmentation de près de 2% absolu du taux d'erreur. L'analyse des erreurs montre que le système reconnait les formes de repli un nombre de fois bien plus important que leurs occurences réelles. Cette surproduction peut soit être due aux variantes de prononciation dans le dictionnaire de prononciation soit aux probabilités assignées par modèle de langage.

En ce qui concerne le modèle de langage, les probabilités attribuées aux formes de repli souffrent de surestimation a l'instar de la forme <UNK> dans le modèle de référence. Cela tient à ce que lors de l'entraînement, le nombre d'occurrences d'une forme de repli fait intervenir les

Forme de repli	*Total des occurrences*	*Taux de reconnaissance*
Nom propre	254	82,3
Nom commun (sg)	49	75,5
Nom géographique	35	82,9
Verbe conjugué (sg)	31	83,9
Verbe conjugué (pl)	26	88,5
Prénom	23	87,0
Nom commun (pl)	21	76,2
Verbe infinitif	15	80,0
Participe passé adjectif (sg)	14	85,7
Adjectif (sg)	11	72,7
Adjectif (pl)	9	66,7
Participe passé (être) (sg)	9	55,6
Participe passé (avoir) (sg)	5	60,0
Participe passé (être) (pl)	5	100,0
Participe passé adjectif (pl)	4	75,0
Adverbe	3	66,7
Participe présent	2	50,0
Nom commun étranger	1	0,0
Total	530	78,7

TAB. 7.6 – Résultats obtenus avec le modèle utilisant le repli lexical. Pour chaque forme de repli, le nombre total d'occurrences et le taux de reconnaissance. Pour les formes déclinées en nombre, le nombre est indiqué par (sg) pour singulier et (pl) pour pluriel.

25	Nom propre	336	Participe passé adjectif (pl)
84	Nom commun (sg)	485	Participe passé adjectif (pl)
111	Nom commun (pl)	615	Participe présent
129	Verbe conjugué (sg)	654	Participe passé (être) (sg)
141	Verbe conjugué (pl)	754	Adverbe
151	Adjectif (sg)	809	Participe passé (être) (pl)
159	Géographique	1233	Participe passé (avoir) (sg)
214	Adjectif (pl)	2487	Mot étranger
266	Infinitif	3143	Cardinal
297	Prénom	5571	Participe passé (avoir) (pl)

TAB. 7.7 – Rang lexical obtenu par les formes de repli. Pour les formes déclinées en nombre, le nombre est indiqué par (sg) pour singulier et (pl) pour pluriel.

occurrences de plusieurs mots. Le tableau 7.7 montre d'ailleurs le rang lexical obtenu à partir du modèle de langage unigramme [4]. Ce tableau montre que les formes de repli ont un rang lexical très élevé puisque le rang varie entre 25 et 5 571 (presque tous sont dans le premier seizième du vocabulaire). En particulier, la forme correspondant aux noms propres est au rang 25 et la forme correspondant aux noms communs singuliers au rang 84, ce qui les classe parmi les mots outils.

Mais la surestimation des probabilités n'est pas suffisante pour expliquer la sur-occurrence des formes de repli, puisque lors de l'expérience interactive, le modèle est le même, et pourtant le phénomène ne s'est pas produit. La méthode d'adaptation *ad hoc* du vocabulaire introduit en moyenne 1 700 mots nouveaux dans le vocabulaire pour un jour donné. Donc, pour une forme de repli, l'introduction des mots nouveaux implique en moyenne 80 variantes de prononciation, alors qu'il n'en a que 3 dans l'expérimentation interactive. 80 n'est qu'une moyenne et ce nombre varie entre 2 et 700. De plus, le nombre d'occurrences des formes de repli dans les transcriptions automatiques montrent que les formes de repli les plus fréquentes sont celles qui ont le plus grand nombre de variantes, avec dans l'ordre décroissant d'apparition dans les transcriptions : les noms propres (700 variantes), noms communs singulier (76 variantes), noms géographiques (68 variantes), prénoms (52 variantes) et noms communs pluriel (43 variantes).

Ainsi la surproduction des formes de repli par le système de reconnaissance automatique de la parole est due au fait qu'une forme de repli possède conjointement une probabilité surestimée par le modèle de langage, et un nombre important de variantes de prononciation. Afin de mesurer l'impact de l'estimation des probabilités associées aux formes de repli, les corpus $A_{28}(j)$ servent à l'estimation des probabilités d'appartenance du mot à sa forme de repli associée $P(w|F(w))$. Ces corpus ont été extraits quotidiennement du Web et utilisés lors de l'adaptation *ad hoc* du vocabulaire. Ils sont décrits au paragraphe 6.1.2. L'expérimentation automatique est alors renouvelée, et les résultats de reconnaissance sont identiques à ceux obtenus avec le système de référence.

[4] le rang lexical s'obtient en classant les mots suivant leurs probabilité unigramme, le rang désigne la place du mot dans ce classement.

Conclusion

La méthode développée dans ce chapitre a pour objectif l'introduction de nouveaux termes dans le vocabulaire d'un modèle de langage *n-gram* de mots sans qu'il soit nécessaire de réentraîner le modèle. Pour cela, des classes particulières nommées formes de repli sont utilisées. Ces formes de repli remplacent lors de l'entraînement les mots nouveaux, qui sont évidement inconnus à ce moment là. Lors de la transcription d'un document, les mots nouveaux sont introduits en tant que variante de prononciation des formes de repli, ce qui permet au modèle de langage de les inclure dans ses distributions.

Cette méthode est expérimentée dans deux cadres d'application. La première expérience est dite interactive. C'est un oracle qui simule la possibilité de compléter manuellement le vocabulaire. Cette expérience montre qu'en annulant la présence de mots hors vocabulaire, il est possible d'obtenir un gain relatif du taux d'erreur de 10,6%. En moyenne, l'annulation du taux de MHV qui était initialement de 1,21%, implique un gain absolu sur le taux de d'erreur de 2,6%, montrant ainsi un rapport moyen de 2,1 entre les deux. Le taux moyen de reconnaissance sur les mots introduits via le repli lexical est de 78,7%. La plupart des erreurs peuvent être attribuées au modèle de langage avec beaucoup de fautes d'accord, et aux conditions acoustiques. Pour les noms propres, les noms de lieu et les prénoms, un tiers des erreurs est due à la difficulté de la phonétisation automatique, en particulier lorsque les noms sont étrangers. Cependant le taux de reconnaissance des mots nouveaux montre que la méthode du repli lexical permet la construction d'un modèle de langage à vocabulaire ouvert et efficace dans ce cadre d'application.

Par contre cette méthode ne semble pas appropriée à l'adaptation automatique du vocabulaire telle qu'elle est définie au chapitre 6. En effet, l'expérience automatique montre que le repli lexical dégrade notablement le taux d'erreur. L'étude des erreurs met en évidence une sur-génération des formes de repli due à un problème d'estimation. Cette sur-génération est due au fait qu'une forme de repli possède conjointement une probabilité surestimé par le modèle de langage, et un nombre important de variantes de prononciation. Une expérience, utilisant un corpus extrait du Web pour affiner l'estimation des probabilités associées, montre que le repli lexical alors ne dégrade pas les performances en terme de taux d'erreur.

Chapitre 8

Adaptation automatique du modèle de langage

Par rapport aux trois paradigmes d'adaptation des modèles de langage présentés à la section 4.5, ce chapitre se situe au niveau du second : celui où le modèle de langage est confronté via le corpus de test à une source de langage inconnue, puisque inobservée lors de son entraînement. Reconstruire un nouveau modèle en incluant cette nouvelle source nécessite de grands corpus d'entraînement, comme ceux décrits au chapitre 4. Ces corpus ne sont pas toujours disponibles pour diverses raisons. Par exemple, pour reprendre les problématiques décrites à la section 2.4, le corpus de test peut contenir des documents d'archive ou, au contraire, des émissions d'actualité très récentes (éventuellement du jour même). Dans ces deux cas, des ressources textuelles au format électronique et en quantité suffisante sont difficilement accessibles. L'adaptation du modèle est alors une solution palliative. Son objectif est d'introduire de l'information relative à cette nouvelle source de langage, tout en conservant les acquis d'une modélisation robuste obtenue à partir de grands corpus d'entraînement.

Dans la première section, la problématique des archives est traitée. Le corpus de test est le corpus d'archives audiovisuelles du projet Echo. Ce corpus couvre près de 50 ans, de 1945 à 1995, et contient un ensemble hétérogène d'émissions télé et radiodiffusées. Une des principales difficultés est d'avoir accès à des données d'entraînement couvrant la période du corpus de test. L'usage de textes provenant de sites Internet est exploré afin d'adapter les modèles de langage à l'époque du document à transcrire.

Dans la seconde section, la problématique traitée est celle du flux entrant, définie au chapitre 1. Les solutions proposées pour adapter le modèle de langage s'appuient sur les méthodes d'adaptation du vocabulaire décrites au chapitre 6, et sur les conclusions du NIST à propos des campagnes *SDR* (voir paragraphe 2.2.4). Deux techniques d'interpolation de modèles de langage sont proposées pour la mise en place d'un modèle de langage tournant (décrit au paragraphe 4.5.3). Deux corpus de test distincts sont d'abord utilisés permettant de montrer l'impact des corpus d'adaptation et de développement sur les résultats de reconnaissance. Puis une expérimentation est effectuée sur des émissions de novembre 2003 diffusées par EuroNews, afin d'évaluer les gains potentiels dans le cas où les données d'adaptation sont pertinentes.

8.1 Transcription automatique d'archives audiovisuelles

Le projet européen ECHO (European CHronicles On-line)[1] avait pour but de fournir un support technologique pour l'archivage numérique de films et d'améliorer l'accessibilité des documents au sein de vastes archives historiques. La transcription automatique de la bande sonore est la première étape d'une analyse automatique du contenu des documents. Le corpus ECHO est constitué d'archives nationales françaises, italiennes, néerlandaises, et suisses. Les données françaises fournies par l'INA ont pour thème la construction européenne, et couvrent la période de 1945 à 1995. Comme le montre une analyse de la transcription automatique, la difficulté n'est pas uniquement de transcrire des documents audiovisuels hétérogènes mais aussi de traiter des données dont les propriétés (qualité d'enregistrement, vocabulaire, style de parole) évoluent au cours du temps. Plusieurs voies ont été explorées pour adapter les modèles statistiques entraînés sur des données récentes aux documents d'archive. L'ensemble de ces expérimentations est décrit dans l'article [Barras *et al.*, 2002], avec en particulier l'adaptation des modèles acoustiques. Seuls les travaux sur l'adaptation des modèles de langage sont repris ici. Le traitement des documents d'archive pose plusieurs problèmes, à commencer par la rareté de données textuelles d'époque permettant une modélisation linguistique adéquate. Des textes des années 1945 à 79 ont été collectés sur un site français d'archives vidéo pour adapter les modèles de langage contemporains.

Dans la suite, le corpus ECHO est décrit, et les résultats de la transcription automatique du corpus avec le système de référence sont analysés. Puis, le corpus de textes extraits du Web est décrit ainsi que son utilisation pour l'adaptation des modèles de langage.

8.1.1 Description du corpus

La partie française du corpus ECHO est constituée de documents audiovisuels d'archive collectés par l'INA. La majeure partie de ces documents avait déjà été rassemblée dans le cadre du projet EURODELPHES[2]. Ce projet avait pour objectif l'édition d'un manuel d'histoire interactif et multimédia sur la construction de la communauté européenne. Depuis ces expérimentations, d'autres documents ont été ajoutés au corpus ECHO provenant de la dernière guerre mondiale, mais pour les expérimentations décrites dans ce chapitre, seul le corpus EURODELPHES a été utilisé.

Le corpus EURODELPHES contient 163 documents au format MPEG-1 et datant des années 1945 à 1995. Ces documents illustrent les étapes de la construction européenne en abordant l'après-guerre ; par exemple les changements de société, les relations Est/Ouest ou la reconstruction. Au total, le corpus représente 9 heures de données audio avec des documents de durées très variables allant de 20 secondes à 20 minutes avec une moyenne de 3 minutes et 20 secondes. Le tableau 8.1 donne une description du corpus par décennie. Il montre que la quantité de données n'est pas uniformément répartie dans le temps et que les décennies les plus représentées sont les années 60 et 70. Certains documents (ou certains segments) sont retirés manuellement du corpus, soit parce qu'ils contiennent de la parole superposée, étrangère ou inaudible, soit pas de parole du tout (chant ou musique). Environ 40 minutes sont ainsi écartées du corpus de test.

[1] http ://pc-erato2.iei.pi.cnr.it/echo/
[2] http ://eurodelphes.ina.fr/

Décennie	*Nombre de documents*	*Durée (minutes)*
1940	13	24
1950	22	53
1960	50	204
1970	31	103
1980	16	64
1990	31	92
Total	163	540

TAB. 8.1 – Nombre de documents et la durée par décennie pour le corpus EURODELPHES.

8.1.2 Analyse des erreurs du système standard

Le système utilisé pour transcrire le corpus de test diffère de celui décrit au chapitre 3 par ses modèles acoustiques. En effet, les données de test ont une bande passante limitée à 6 kHz, alors que les modèles acoustiques ont été entraînés sur des données ayant une bande passante de 8 kHz. Les modèles ont donc été au préalable ré-estimés en réduisant la bande passante de leurs corpus d'entraînement à 6 kHz. Une réduction relative de 3% du taux d'erreur sur les mots a été ainsi obtenue. Ce sont ces modèles acoustiques qui sont utilisés dans la suite des expérimentations sur le corpus ECHO.

Les résultats sont donnés par décennie dans le tableau 8.2. Le taux de mots hors vocabulaire augmente avec l'âge des documents et il varie suivant les décennies entre 0,6% pour les années 90 et 1,5% pour les années 40, avec une valeur moyenne de 0,9%. Ces valeurs sont surprenantes puisqu'elles sont inférieures à celle obtenue généralement sur un corpus récent d'émissions d'actualité. Ceci est probablement dû au fait que le corpus est construit autour du thème commun de la construction européenne. Une étude des mots hors vocabulaire montre que 75% des formes hors vocabulaire n'apparaissent qu'une seule fois, que 37% de leurs occurrences sont des noms propres, 28% sont des formes conjuguées de verbes déjà dans le vocabulaire, et que 11% sont des variantes en genre et en nombre de mots du vocabulaire. Pour les MHV, la part des noms propres augmente avec l'âge des documents puisqu'ils représentent la moitié des MHV dans les années 40 et moins de 20% dans les années 80 et 90. La perplexité est de 99 sur l'ensemble du corpus. Cependant elle augmente nettement avec l'âge des données montrant l'impact de l'époque des données d'entraînement du modèle de langage. Le modèle de langage a été entraîné avec des textes qui sont principalement des années 90, et d'ailleurs, la perplexité des données de test de cette période est de 81, alors que la perplexité pour les années 40 est de 133.

L'évolution du taux d'erreur au cours des décennies est différente. Les années les plus difficiles à traiter sont les années 70 et 80 où le taux d'erreur avoisine les 40%. Ce résultat peut être attribué à la répartition non uniforme des documents selon les décennies. L'étude des erreurs du système de reconnaissance pour les documents conduisant aux taux d'erreur les plus élevés tend à le montrer. Pour les années 70, les segments de parole se répartissent de la manière suivante : 32% des mots sont contenus dans 5 des 28 documents sur lesquels le taux d'erreur est de 56%. Beaucoup de segments de ces émissions contiennent de la parole bruitée ou superposée. Certains

Décennie	*# Mots*	*%MHV*	*Perplexité*	*WER*
1940	3148	1,5	133	34,1
1950	8029	1,3	120	33,1
1960	28365	1,1	109	36,7
1970	15332	0,5	92	41,0
1980	9667	0,8	94	39,7
1990	16807	0,6	81	28,6
Total	81348	0,9	99	**35,7**

TAB. 8.2 – Nombre de mots, taux de mots hors vocabulaire, perplexité et taux d'erreur obtenus avec le système de référence modifié.

segments ont été purement et simplement écartés par le partitionneur et n'ont donc pas été reconnus. Sur les autres émissions de la décennie, le taux d'erreur est de 33%. Pour les années 80 le contraste est encore plus marqué, puisque 43% des mots sont contenus dans un seul document sur les 16 de la décennie, et sur lequel le taux d'erreur est de 51%. Le principal locuteur de cette émission est un chômeur dépressif qui s'exprime de manière sous-articulée. Sur le reste de la décennie, le taux d'erreur n'est plus que de 31%.

La transcription des émissions des années 40 et 50 pose moins de problèmes. Une des raisons est que dans les documents les plus anciens, les conditions d'enregistrement étaient plus contraignantes. Les locuteurs y ont un discours préparé. De plus, le débit de parole s'accélère avec le temps, puisqu'il était de 160 mots par minute dans les années 40, puis de 170-175 dans les années 50 à 70 et de 185 mots par minute dans les années 80 et 90. Ce style de parole plus clair semble compenser les difficultés rencontrées du fait de l'augmentation du taux de MHV et de la perplexité avec le temps.

8.1.3 Adaptation des modèles de langage

La couverture lexicale du vocabulaire de référence est en moyenne meilleure sur le corpus de test que ce qui est usuellement observé sur un corpus récent d'émission d'actualité. Le vocabulaire de référence est par conséquent conservé. Afin d'adapter le ML, des données d'époque sont nécessaires. Or, peu de textes des années 40 à 80 sont disponibles dans un format électronique. Sur le site des archives Gaumont[3], les notices documentaires des productions audio et vidéo de Gaumont de 1900 à nos jours sont accessibles via un moteur de recherche plein texte. Ces données sont utilisées pour estimer un modèle de langage « d'époque ». Ce modèle permet ensuite d'adapter le modèle de langage de référence.

Corpus extrait du Web

Les notices documentaires correspondant aux périodes du corpus ECHO ont été téléchargées à partir du site des archives Gaumont. Comme le montre la figure 8.1, le contenu de ces notices est

[3]http ://www.newsreels.gaumont.com

Document :	105 / 2377
No Référence :	5300GNU09956
Année :	1953
Texte :	Course d'autos.
Origine :	GAUMONT
Catégorie :	NON UTILISE

Document :	1053 / 2377
No Référence :	5312EJ 44683
Année :	1953
Texte :	Sous marin monopole, essai de plongée. AFFICHETTE. Techniques. Sous marin individuel pour " homme grenouille ".
COMMENTAIRE :	Les hommes grenouille américains expérimentent un nouveau dispositif de propulsion destiné à leur éviter toute fatigue dans leur déplacement sous-marin. Véritable soucoupe flottante, l'engin leur assure une alimentation complémentaire en air respirable et leur permet de naviguer à grande vitesse jusqu'à des profondeurs de 100 mètres. Un moteur à air comprimé agissant par réaction sur l'eau donne la dernière touche ultra moderne au taxi des profondeurs qui, espérons-le, fera bientôt le bonheur des amateurs de pêche sous-marine.
Origine :	ECLAIR
Catégorie :	JOURNAUX
Métrage :	17 mètres, 0 mn 0 s.

FIG. 8.1 – Deux exemples de notices : une première notice concise, et une seconde plus riche.

très variable. Le champ « Texte » est visiblement issu d'un formulaire qui a été déstructuré. La première étape du traitement consiste à retrouver une structure permettant de trier l'information utile de celle qui par exemple spécifie le support de conservation du document. Ensuite, lors de la seconde étape, les textes sont normalisés comme cela est décrit à la section 4.4.2. Quelques problèmes spécifiques à la structure de ces notices se posent. Les noms propres sont écrits en majuscules et leur traitement ne doit pas interférer avec le traitement des sigles et acronymes, ou de la première lettre de chaque phrase. Mais le problème le plus délicat est l'haplologie du point qui sert pêle-mêle de séparateur de phrases, de champs (texte, commentaire, ...), d'énumération, de sigles ou acronymes (pas de manière systématique), et remplace parfois la virgule pour les nombres réels. Puisque le vocabulaire de référence est conservé, les choix de normalisation ont pour objectif de réduire le nombre de mots hors vocabulaire.

Le corpus d'entraînement obtenu (corpus Gaumont) contient 2,7 million de mots. Sa répartition sur les années 45 à 80 est représentée dans la figure 8.2. Les textes ne sont pas uniformément répartis sur la période, et une majorité d'entre eux date des années 50 et 60 : 1,4 million pour les années 50, et un million pour les années 60. Cette répartition est proche de celle des documents audio dans le corpus de test. Le taux de mots hors vocabulaire mesuré avec le vocabulaire de référence sur le corpus Gaumont est de 4%.

Modèle de langage d'époque

Étant donnée la quantité de données et leurs répartitions, un seul modèle de langage a été estimé à partir des données d'entraînement décrites au paragraphe précédent. Ce modèle quadrigramme contient 190 149 bigrammes, 252 018 trigrammes et 191 667 quadrigrammes[4]. La perplexité obtenue avec ce modèle sur les transcriptions de référence est de 333 sur l'ensemble du corpus, et décroît de 374 pour les années 90 à 274 pour les années 40. L'ensemble des résultats est donné dans le tableau 8.3.

Adaptation *vs* époque

L'adaptation du modèle de référence est réalisée par son interpolation linéaire avec le modèle d'époque. Cette interpolation, pour tout mot w du vocabulaire et son historique h, s'écrit comme il suit :

$$P_\lambda(w|h) = \lambda P_{epoque} + (1 - \lambda) P_{ref}(w|h) \tag{8.1}$$

où P_{epoque} désigne la probabilité assignée par le modèle d'époque, $P_{ref}(w|h)$ celle assignée par le modèle de référence et $P_\lambda(w|h)$ la probabilité assignée par le modèle interpolé résultant. L'influence du coefficient d'interpolation λ est étudiée en lui attribuant les valeurs suivantes $\lambda = \{ 0,1 \ , \ 0,2 \ , \ 0,3 \ \}$. Un modèle adapté M_λ est associé à chacune de ces valeurs : $M_{0,1}$, $M_{0,2}$, et $M_{0,3}$. Le modèle $M_{0,0}$ est le modèle de référence et $M_{1,0}$ le modèle d'époque.

Avec chacun de ces modèles, la perplexité des données de test est mesurée par décennie. Les résultats sont rassemblés dans le tableau 8.3. Pour chaque décennie, le « meilleur » modèle désigne le modèle qui permet d'obtenir la perplexité la plus faible. La lecture par ligne du tableau montre que pour les années 40, la perplexité est une fonction décroissante du coefficient d'interpolation, et le meilleur modèle est $M_{0,3}$. Ce modèle est celui qui donne le plus d'importance au modèle d'époque. Par contre, pour les années 90, la perplexité est une fonction croissante du coefficient d'interpolation, et le meilleur modèle est le modèle de référence. La figure 8.3 représente l'évolution par décennie de la perplexité, obtenue avec le modèle de référence et avec le meilleur modèle. La figure montre que le gain en perplexité obtenu avec l'adaptation est maximale pour les années 40, puis qu'il ne cesse de décroître au cours des décennies : 9%, 6%, 3%, 1% des années 40 à 70. Pour les années 80 et 90, le meilleur modèle est le modèle de référence.

Adaptation non supervisée

Dans un cadre opérationnel, il est possible de procéder à une adaptation non supervisée en estimant le coefficient d'interpolation λ_u via l'algorithme E.M. sur les transcriptions automatiques

[4] Comme lors de la construction du modèle de langage de référence (voir la section 4.4), des seuils sur le nombres d'occurences des ngrammes sont appliqués (*cut-off*).

Décennie	$M_{0,0}$	$M_{0,1}$	$M_{0,2}$	$M_{0,3}$	M_{epoque}	λ_u	*WER Réf*	*WER Ad*
1940	133	125	122	**121**	274	0,30	33,1	34,1
1950	120	114	**113**	114	293	0,16	32,9	33,1
1960	109	**106**	**106**	107	318	0,24	36,9	36,7
1970	92	**91**	92	94	335	0,08	41,1	41,0
1980	**94**	**94**	95	98	366	0,06	39,4	39,7
1990	**81**	82	83	86	374	0,04	28,6	28,6
Total	99	**97**	98	100	333	0,16	35,7	35,7

TAB. 8.3 – Perplexité par décennie estimée sur les données de test avec le modèle standard, le modèle d'époque et les modèles adaptés, puis λ_u le coefficient d'interpolation estimé sur les transcriptions automatiques effectuées avec le système de référence. Les perplexités en gras sont les plus faibles pour une décennie donnée et désigne le meilleur modèle. Les taux d'erreur de reconnaissance ont été obtenus avec le modèle de référence (WER Réf) et le meilleur modèle (WER Ad).

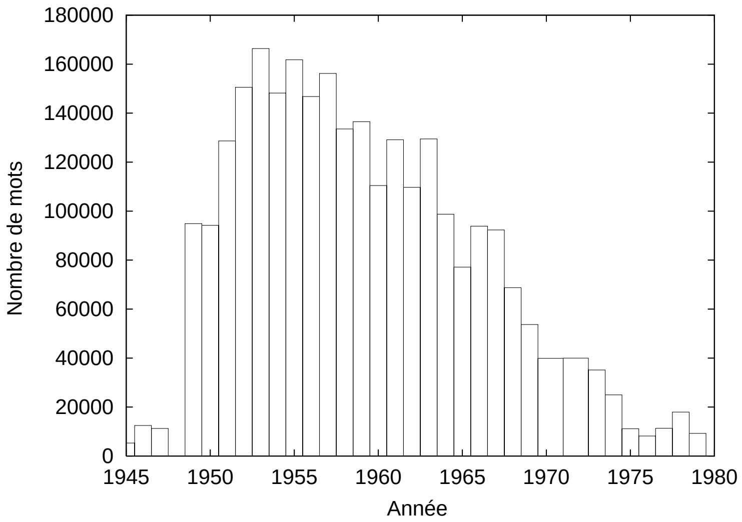


FIG. 8.2 – Répartition des textes extraits du Web suivant les années.

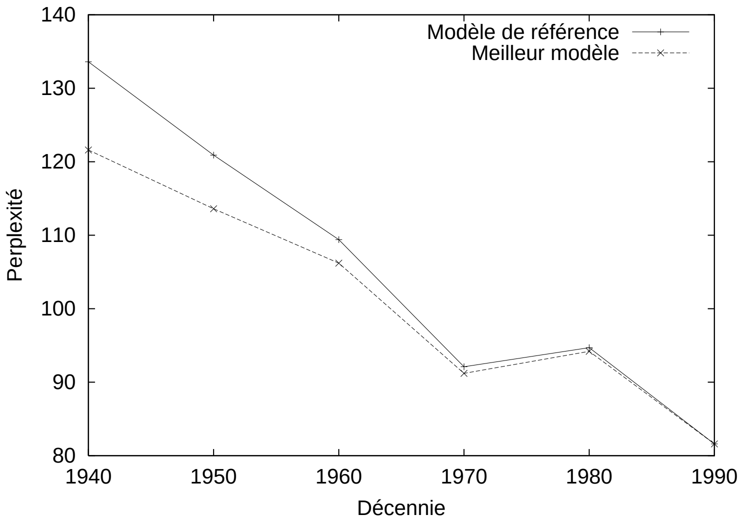


FIG. 8.3 – Perplexité obtenue avec le modèle de référence et avec le meilleur modèle.

obtenues avec le système de référence. Les valeurs obtenues sont également rassemblées dans le tableau 8.3. En arrondissant λ_u aux valeurs $\{0,\ 0,1,\ 0,2,\ 0,3\}$, le coefficient λ_u correspond effectivement au coefficient d'interpolation du meilleur modèle. Ce dernier a été utilisé pour les expériences de reconnaissance dont les résultats sont donnés dans la dernière colonne du tableau 8.3. Sur l'ensemble du corpus de test, le WER est le même avec ou sans adaptation. Deux raisons peuvent être évoquées. La principale est que le gain en perplexité est maximum pour les années 40, alors que pour les autres décennies, il semble insuffisant pour garantir un gain sur le taux d'erreur[5]. Le seul gain significatif est observé pour les années 40 avec le modèle $M_{0,2}$. Le taux d'erreur est alors de 32,9%.

[5]Ceci semble vérifier la règle énoncée à la section 4.3.2 sur l'impact des gains en perplexité sur le taux d'erreur

8.2 Transcription automatique d'émission d'actualité

Au chapitre 6, deux méthodes d'adaptation automatique du vocabulaire ont été introduites. Toutes deux modélisent à partir de textes extraits du Web le contenu lexical de l'actualité et permettent une réduction significative du taux de MHV sur un corpus de développement extrait lui aussi d'un site Internet. L'étape suivante consiste à construire un modèle de langage capable d'assigner une probabilité à toute suite de mots extraite du nouveau vocabulaire. Les quantités de données sont insuffisantes pour reconstruire un modèle de langage aussi complet que le modèle de référence M_{ref}. Il est par contre possible d'estimer un modèle de langage $M_a(j)$ pour chaque jour j à partir d'un corpus d'entraînement constitué de textes du jour collectés sur le Web. Le modèle de référence est adapté en l'interpolant avec le modèle $M_a(j)$. Lorsque l'adaptation d'un modèle de langage ne se fait pas à vocabulaire constant, certaines modifications doivent être apportées pour assurer la cohérence de l'interpolation, comme celle décrite par exemple au paragraphe 4.5.3. Deux types d'interpolation sont proposés qui permettent de maintenir le vocabulaire dans les limites des 65 535 mots. Ces méthodes d'adaptations sont expérimentées sur deux corpus de test distincts : l'un est constitué de journaux télévisés de la chaîne TF1 et l'autre de journaux radiodiffusés par la chaîne RFI.

Ces deux corpus de test diffèrent à plus d'un titre, d'abord parce que chaque diffuseur a un point du vue éditorial qui lui est propre (l'actualité internationale est, par exemple, plus détaillée et plus variée dans les journaux de RFI), puis parce que la forme et le média sont différents. En particulier, les reportages sont moins nombreux dans les émissions de radio, ce qui réduit la proportion de parole spontanée au profit d'un discours journalistique plus proche du langage de la presse écrite. Or, les textes servant à estimer le modèle $M_a(j)$ sont eux aussi issus de la presse écrite. De plus, dans les corpus décrits à la section 6.1, un corpus particulier est constitué des résumés quotidiens de l'actualité, disponibles sur le site de la radio RFI. Ce corpus a donc été choisi comme corpus de développement pour la méthode d'adaptation vectorielle. Ainsi, pour transcrire les émissions de RFI une partie des données d'entraînement et de développement proviennent du site Web de RFI, ce qui n'est pas le cas lorsque les journaux télévisés de TF1 sont transcrits.

8.2.1 Corpus de test et d'adaptation

Les corpus d'adaptation sont construits à partir des corpus décrits à la section 6.1. La séparation entre test, développement était alors justifiée pour évaluer les 2 méthodes d'adaptation du vocabulaire. Désormais, tous les textes disponibles sont utilisés pour entraîner les modèles d'adaptation. Pour chaque jour j, le corpus d'entraînement de $M_a(j)$ est $Adapt_{28}(j) = A_{28}(j) \cup D_{28}(j) \cup T_{28}(j)$, il réunit ainsi tout les textes disponibles le jour j et les 27 derniers jours. Le corpus de développement nécessaire à la méthode d'adaptation vectorielle du vocabulaire est constitué à partir des textes du site de RFI : $T_{14}(j)$ (au chapitre 6, ce corpus servait de test).

Deux corpus de test distincts issus du mois de janvier 2002 sont utilisés. Le premier est le corpus de test déjà utilisé au paragraphe 5.1.2. Il rassemble les journaux télévisés de 20h, recueillis tous les jours du mois de janvier 2002 sur le site internet de la chaîne de télévision française

TF1 au format RealVideo. Afin d'obtenir un échantillon représentatif sur le mois de janvier, le premier quart d'heure de tous les journaux, à l'exception des samedis et dimanches, a été transcrit manuellement avec le logiciel d'aide à l'annotation de signaux de parole *Transcriber* [Barras *et al.*, 2001]. Trois journaux, répartis sur le mois, ont été intégralement transcrits (les 3, 15 et 30 janvier). Le premier quart d'heure d'un journal télévisé représente en moyenne 2 850 mots et un journal complet 7 930 mots. Dans sa totalité, le corpus contient 77 541 mots pour 7 h 15 de signal. Le second corpus de test est constitué à partir de la tranche horaire consacrée à l'actualité et diffusée de 7 h à 8 h sur RFI, pour les 3, 15 et 29 janvier 2002. Il contient au total 33 506 mots pour 3 heures de signal. Cette tranche horaire a un contenu similaire à un journal télévisé avec des interviews et des reportages mais l'actualité internationale est plus développée. Les taux d'erreurs ont été calculés avec les outils d'évaluation du NIST[6].

8.2.2 Interpolation linéaire

L'adaptation du modèle de langage est effectuée pour chaque jour par l'interpolation linéaire du modèle de référence avec le modèle d'adaptation $M_a(j)$. Or le vocabulaire de ce dernier peut être différent de celui du modèle de référence. En particulier, pour construire $M_a(j)$, deux vocabulaires adaptés sont expérimentés. Il est donc nécessaire d'interpoler deux modèles de langage qui n'ont pas le même vocabulaire.

L'interpolation linéaire s'écrit dans le cas présent :

$$P_{interpol}(w_i|h_i) = \lambda P_a(w_i|h_i) + (1-\lambda) P_{ref}(w_i|h_i) \tag{8.2}$$

Les probabilités $P_{ref}(w_i|h_i)$ et $P_a(w_i|h_i)$ sont les probabilités assignées par chacun des modèles de langage faisant intervenir le repli à l'ordre inférieur si nécessaire. Pour un modèle *n-gram*, on a $h_i = w_{i-n+1}, ...w_{i-1}$.

Dans la section 8.1, le vocabulaire de référence était imposé pour la construction du modèle d'adaptation (le modèle d'époque), puisque le corpus de test ne nécessitait pas de modification du vocabulaire. L'objectif ici est non seulement d'adapter le modèle de référence aux évolutions thématiques de l'actualité qui se traduisent par des changements lexicaux et linguistiques, mais aussi d'introduire les mots nouveaux issus de l'adaptation du vocabulaire dans le modèle de langage. Dans l'équation 8.2, les deux modèles interpolés n'ont donc pas le même vocabulaire. Des précautions doivent être prises pour appliquer cette équation, et pour conserver un vocabulaire de taille inférieure à 65 535 mots.

Pour respecter la contrainte sur la taille du vocabulaire, le vocabulaire du modèle adapté est choisi comme vocabulaire du modèle interpolé. Le ML de référence est restreint à ce vocabulaire en supprimant tous les ngrammes qui contiennent au moins un mot n'appartenant pas au vocabulaire commun. La normalisation des probabilités n'est plus strictement respectée dans tous les cas :

$$\sum_{w \in V_{commun}} P_{ref}(w|h) \leq 1$$

[6]http ://www.nist.gov/speech/

Pour appliquer l'équation 8.2, un ensemble de règles similaires à celle décrite au paragraphe 4.5.3 est appliqué au modèle de référence. Elles sont décrites ci-dessous pour un modèle trigramme :

$$\begin{aligned} P_{ref}(x|w_{i-2}w_{i-1}) &= 0 \\ P_{ref}(w_i|xw_{i-1}) &= P_{ref}(w_i|w_{i-1}) \\ P_{ref}(w_i|w_{i-2}x) &= P_{ref}(w_i) \end{aligned} \tag{8.3}$$

où w_{i-2}, w_{i-1}, w_i désignent des mots appartenant au vocabulaire de référence, et x un mot n'y appartenant pas. Le coefficient d'interpolation λ a été fixé empiriquement à $0,1$ après quelques tests de transcription automatique.

8.2.3 Interpolation sélective

Lors de l'interpolation linéaire, les probabilités peuvent être obtenues par repli. En particulier, si un historique contient un mot nouveau introduit grâce à l'adaptation du vocabulaire, la distribution qui lui est associée est modifiée par le modèle de référence qui ne l'a pourtant jamais rencontré lors de son entraînement. Ceci est également vrai dans l'autre sens, puisque le corpus d'entraînement du modèle $M_a(j)$ est réduit, beaucoup de ngrammes n'ont pas été observés lors de son entraînement et malgré cela leur probabilité assignée par le modèle de référence est modifiée. Une nouvelle méthode d'interpolation sélective est proposée, qui tient compte des observations de chacun des modèles au niveau des historiques pour construire le modèle P_{sel} :

$$P_{sel}(w_i|h_i) = \begin{cases} P_{ref}(w_i|h_i) & \text{, si } N_a(h_i) = 0 \\ P_a(w_i|h_i) & \text{, si } N_{ref}(h_i) = 0 \\ \lambda P_a(w_i|h_i) + (1-\lambda) P_{ref}(w_i|h_i) & \text{, sinon} \end{cases} \tag{8.4}$$

où $N_{ref}(h_i)$ désigne le nombre d'observations de l'historique h_i lors de l'entraînement du modèle de référence et de même pour $N_a(h_i)$ dans le cas du modèle $M_a(j)$. Les deux valeurs extrêmes de λ impliquent un comportement particulier du modèle M_{sel}. Lorsque $\lambda = 0$, le modèle M_{sel} est une fusion des deux modèles M_{ref} et M_a qui donne la priorité au modèle de référence : lorsqu'un ngramme a été observé lors de l'entraînement des deux modèles, seule la probabilité du modèle de référence est prise en compte. Réciproquement, dans le cas $\lambda = 1$, la priorité est donnée au modèle $M_a(j)$. Le coefficient d'interpolation λ a été fixé empiriquement à $0,1$ après quelques tests de transcription automatique.

8.2.4 Expérimentation sur les émissions radiodiffusées

Pour chaque jour, trois choix sont expérimentés pour le vocabulaire du modèle d'adaptation : le vocabulaire de référence, le vocabulaire adapté selon la méthode *ad hoc* et le vocabulaire vectoriel. Pour chacun de ces trois choix, un premier décodage complet en trois passes est effectué conformément à ce qui a été décrit au paragraphe 3.1.5. Le modèle de langage utilisé est le modèle adapté via l'interpolation linéaire (équation 8.3). Les résultats de reconnaissance

Jour	*#mots*	*Réf*	*Ad hoc*	*Vect*
8	11329	1,27	0,99	0,88
15	11023	1,05	0,88	0,88
29	11154	1,03	0,91	0,96
Total	33506	1,12	0,92	0,91

TAB. 8.4 – Taux de mots hors vocabulaire obtenus sur le corpus de test RFI, avec le vocabulaire de référence, puis les deux vocabulaires adaptés, le premier selon la méthode *Ad hoc* et le second selon la méthode vectorielle (*Vect*).

	Vocabulaire		
Jour	8	15	29
8	0,88	0,94	1,00
15	1,12	0,88	0,87
29	0,99	0,99	0,96

TAB. 8.5 – Taux de mots hors vocabulaire obtenus avec les vocabulaires adaptés pour chaque jour via la méthode vectorielle, et calculés sur les transcriptions du corpus de test RFI.

relatifs à l'interpolation sélective sont obtenus en recalculant le treillis de la troisième passe (rescoring). Des expérimentations ont été menées, effectuant le processus de décodage complet pour chaque modèle. Outre le fait, que le temps de calcul en est augmenté, les résultats ne sont pas significativement différents de ceux obtenus en recalculant uniquement le treillis.

Couverture lexicale

Le tableau 8.4 rassemble les taux de MHV observés sur le corpus de test obtenus avec le vocabulaire de référence, puis les deux vocabulaires adaptés, le premier selon la méthode *ad hoc*, et le second selon la méthode vectorielle (*Vect*). Le gain relatif apporté par l'adaptation du vocabulaire sur le taux de MHV est de 18% en moyenne quelle que soit la méthode utilisée. Ces résultats sont donc différents de ceux présentés au chapitre 6, puisque les gains sont divisés par plus d'un facteur 2. Cette différence peut être imputée à la différence entre le langage écrit et le langage parlé, aux erreurs de normalisation et surtout au fait que les corpus d'entraînement utilisés pour adapter les vocabulaires couvrent une actualité bien plus large qu'une émission de radio ou de télévision.

La pérennité de l'adaptation vectorielle peut être étudiée grâce aux résultats du tableau 8.5. Le meilleur vocabulaire est presque toujours celui adapté pour le jour même (sauf pour la transcription du 15 janvier où la différence de performance avec le vocabulaire adapté pour le 29 est peu sensible). Et utiliser un autre que celui du jour même peut dégrader la couverture lexicale de manière significative, avec par exemple pour la transcription du 15 janvier une perte de couverture lexicale de près d'un quart avec le vocabulaire du 8. La même étude avec la méthode d'adaptation *ad hoc* donne des résultats différents avec des taux de mots hors vocabulaire bien plus stable (le tableau correspondant est donné en annexe 8.12).

	Modèle de langage		
Jour	Réf	Linéaire	Sélective
8	155	147	141
15	128	120	117
29	125	120	117
Total	136	129	125

TAB. 8.6 – Perplexité des transcriptions de référence, calculées avec le modèle de référence, et deux modèles adaptés via l'interpolation linéaire et sélective et contruits avec le même vocabulaire vectoriel.

Jour	*#mots*	*Réf*	*Vocabulaire de référence*		*Vocabulaire adhoc*		*Vocabulaire vectoriel*	
			linéaire	sélective	linéaire	sélective	linéaire	sélective
8	11329	22,4	22,1	22,1	21,5	21,3	21,6	21,2
15	11023	20,2	19,7	19,8	19,3	19,3	19,2	19,2
29	11154	17,7	17,6	17,5	17,5	17,6	17,5	17,4
Total	33506	20,1	19,8	19,8	19,4	19,4	19,4	19,3

TAB. 8.7 – Résultats de reconnaissance obtenus avec le modèle de référence et les modèles adaptés selon le vocabulaire utilisé soit par interpolation linéaire soit interpolation linéaire modifiée.

Résultats de reconnaissance

Afin de dissocier l'impact du vocabulaire de celui de l'adaptation des ML, sept expériences ont été menées dont les résultats sont donnés dans le tableau 8.7. La première expérience utilise le ML de référence. Puis, pour chaque vocabulaire (le vocabulaire de référence et les deux vocabulaires adaptés), le ML de référence est adapté dans un premier temps par son interpolation linéaire avec $M_a(j)$ et dans un second temps par son interpolation sélective avec $M_a(j)$.

Le tableau 8.6 rassemble les perplexités des transcriptions de référence, calculées avec différents modèles de langage : le modèle de référence, puis deux modèles adaptés via l'interpolation linéaire et sélective et construits avec le même vocabulaire vectoriel. Suivant la méthode d'adaptation du vocabulaire, vectoriel ou *ad hoc*, les résultats obtenus diffèrent d'environ 1%. Cette différence est uniquement due à la différence de couverture lexicale puisque les textes d'entraînement et les coefficients d'interpolation sont les mêmes.

L'adaptation des modèles de langage permet donc un gain en perplexité de 5% avec l'interpolation linéaire et, 9% avec l'interpolation sélective. Entre les deux méthodes d'adaptation, la différence en perplexité est de 4% en faveur de l'interpolation sélective.

Les meilleurs résultats sont toujours obtenus avec la configuration suivante : pour le vocabulaire, une adaptation vectorielle, et pour le modèle de langage, une adaptation utilisant l'interpolation sélective. Cependant, à vocabulaire constant, les résultats diffèrent peu suivant l'interpolation choisie pour adapter les modèles de langage. La seule différence sensible s'observe pour l'émission du 8 janvier et avec le vocabulaire adapté selon la méthode vectorielle. Par contre, à méthode d'interpolation constante, l'adaptation du vocabulaire a un impact significatif sur les performances.

La comparaison des tableaux 8.7 et 8.4 montre que lorsque le taux de mots hors vocabulaire augmente le taux d'erreur augmente également. Les gains relatifs en terme de taux d'erreur sont pour le 8, 15 et 29 janvier de 5,4%, 5,0% et 1,7% respectivement, avec une moyenne de 4%. Parallèlement, en terme de taux de mots hors vocabulaire les gains relatifs suivent la même évolution : 31%, 16% et 7%. Les gains absolus sur le taux de MHV sont au total avec le vocabulaire vectoriel de 0,2%. Avec ce vocabulaire et un modèle de langage adapté via l'interpolation sélective, le ratio entre la réduction du taux d'erreur est la réduction du taux de MHV est de 4. Donc, si en moyenne, un mot hors vocabulaire apparaissant dans le document à transcrire entraîne 1,5 à 2 erreurs de reconnaissance, alors l'adaptation du modèle de langage implique une amélioration de la transcription au-delà de ce que le gain en couverture lexicale peut laisser présager.

Transcription des mots nouveaux

Examinons les résultats de la transcription automatique sur les mots qui, parmi ceux ajoutés par l'adaptation vectorielle, apparaissent dans les documents à transcrire. Pour cela, le vocabulaire adapté est séparé en deux catégories : les mots déjà présents dans le vocabulaire de référence et les mots nouveaux. De cette dernière catégorie, ne sont conservés que les mots qui apparaissent dans les transcriptions de référence, et pour chacun, il est vérifié si le système l'a transcrit correctement ou pas.

En moyenne, l'algorithme vectoriel modifie 17% du vocabulaire soit environ 11 400 mots. Mais ces mots nouveaux correspondent seulement à 121 occurrences pour 81 formes (sur plus de 33 000) dans les transcriptions de référence . Sur ces 121 occurrences, 70 sont correctement transcrites, soit 60%. Le tableau 8.8 résume en pourcentage la répartition des mots nouveaux apparaissant dans les transcriptions de référence. Les catégories « Nom propre étranger » et « Nom commun étranger » désignent des mots d'origine étrangère et qui ne suivent pas les règles usuelles de prononciation française. Par exemple, « Duhalde » est classé dans les noms propres étrangers contrairement à « Poutine ».

Les erreurs concernant les noms propres étrangers peuvent presque toutes être imputées à la phonétisation automatique. Par exemple, le mot « Duhalde » est transcrit par /dhald/ au lieu de /dualde/, ou « Enron » par /ArO/ alors qu'il est prononcé /Enrxn/ ou /Enrcn/. Pour les noms propres en général et les noms propres étranger en particulier, les prononciations observées sont souvent variables : le nom « Sirven » peut se prononcer /sirvEn/ ou /sirvI/, ou par exemple le prénom « Yitzhak » (Rabin) se prononce /itsak/ ou /izak/. Le problème n'est donc pas uniquement de pouvoir générer une phonétisation correcte mais également les variantes possibles. Les quelques erreurs qui ne sont pas directement dues à la phonétisation le sont indirectement puisqu'elles résultent de la présence dans le voisinage de mots dont la phonétisation est erronée. L'exemple le plus fréquent est la séquence « prénom nom » comme « Eduardo Duhalde ».

Pour les autres catégories, la phonétisation automatique n'est pas remise en cause, et les erreurs sont dues à la présence d'un mot hors vocabulaire dans le voisinage, à des conditions acoustiques difficiles pouvant entraîner des erreurs de segmentation, ou encore à des fautes d'accord imputables au modèle de langage.

	% des occurences	%Correctement transcrits
Nom propre étranger	53	52
Nom propre	21	62
Nom commun étranger	2	100
Nom commun	11	77
Adjectif	5	50
Verbe conjugué	8	60

TAB. 8.8 – Répartition des mots nouveaux présents dans les transcriptions de référence, et pour chaque type la proportion correctement transcrite.

8.2.5 Expérimentation sur les émissions télédiffusées

Les expériences décrites dans ce paragraphe sont menées de manière identique à celles du paragraphe 8.2.4. Seul le corpus de test est changé, puisqu'il s'agit désormais de journaux télévisés de TF1. Des deux corpus de test, celui-ci est *a priori* le plus difficile. En effet, pour le corpus de test radiodiffusé, une partie des données d'adaptation utilisées provient du site Web du diffuseur, ce qui n'est pas le cas ici.

Couverture lexicale

Sur l'ensemble du mois de janvier, les taux de mots hors vocabulaire obtenus avec le vocabulaire de référence et les deux vocabulaires adaptés suivant les méthodes *ad hoc* et vectorielle sont respectivement de 1,13% 0,99% 1,15%. En ce qui concerne l'adaptation vectorielle du vocabulaire, ces résultats montrent la dépendance de la méthode au corpus de développement utilisé : lorsque le corpus de développement est lié à la cible (comme au paragraphe 8.2.4), la méthode permet d'obtenir un gain relatif de 18%, alors que dans le cas contraire, elle dégrade légèrement les résultats. Les résultats détaillés par jour sont donnés en annexe dans le tableau 8.13. La comparaison quotidienne des taux de MHV obtenus avec le vocabulaire de référence et le vocabulaire vectoriel montre des différences relatives qui varient entre -38% et 22%, suivant l'adéquation entre le corpus de développement et le corpus de test.

Par contre la méthode d'adaptation *ad hoc* maintient à peu près ses performances avec un gain relatif de 16%. Au niveau d'une émission, l'algorithme *ad hoc* dégrade les performances qu'une seule fois de 4% relatif, par contre le gain maximum relatif est de 26%. La méthode *ad hoc* permet donc une amélioration significative de la couverture lexicale, et contrairement au expérimentation du paragraphe 8.2.4, elle est plus performante que l'algorithme vectoriel puisqu'elle ne nécessite pas de données de développement.

Résultats de reconnaissance

Sur l'ensemble du corpus de test, les résultats de reconnaissance et les perplexités calculés sur les transcriptions de référence sont donnés dans le tableau 8.9. Quel que soit le vocabulaire

	Réf	*Vocabulaire de référence*		*Vocabulaire vectoriel*		*Vocabulaire ad hoc*	
		linéaire	sélective	linéaire	sélective	linéaire	sélective
% de MHV	1,13	1,13		1,15		0,99	
Perplexité	127	119	118	120	119	120	119
WER	24,2	24,1	24,0	23,9	23,8	23,8	23,7

Tab. 8.9 – Résultats de reconnaissance obtenus sur le corpus de test télédiffusé, en faisant varier le vocabulaire (référence, adaptation *ad hoc* et vectoriel) et la technique d'adaptation des modèles de langage (interpolation linéaire et sélective). Les résultats du modèle de langage de référence sont également indiqué (*Réf*).

utilisé ou la méthode d'adaptation choisie, le gain en perplexité est à peu près identique autour de 7%.

Les résultats détaillés pour chaque émission sont donnés en annexe dans le tableau 8.14 pour l'ensemble du corpus de test. En terme de taux d'erreur, quel que soit le vocabulaire adapté choisi, l'adaptation des modèles de langage permet une réduction significative. Par contre, la différence entre les méthodes est très faible bien qu'en faveur de l'interpolation sélective. Les résultats de reconnaissance reflètent les résultats observés sur la couverture lexicale, puisque les meilleurs résultats sont obtenus en utilisant le vocabulaire adapté selon la méthode *ad hoc*. Cependant, les gains sont nettement inférieurs à ceux obtenus sur le corpus de test radiodiffusé : 2,1% en relatif avec la méthode *ad hoc* et 1,7% avec la méthode vectorielle (l'adaptation utilisant l'interpolation sélective). Notons qu'à perplexité constante (égale à 119), le taux d'erreur peut varier de 0,4% absolu (comparaison de la troisième et dernière colonne du tableau 8.14). Il est difficile de distinguer une corrélation entre les trois grandeurs que sont la couverture lexicale, la perplexité, et le taux de d'erreur. Par exemple, la comparaison entre d'une part le modèle adapté par interpolation linéaire et construit avec le vocabulaire de référence, et d'autre part, le modèle adapté par interpolation sélective et construit avec le vocabulaire vectoriel, montre que malgré une perplexité identique, une différence de 0,3% absolu sur le taux d'erreur est observée en faveur du modèle utilisant le vocabulaire vectoriel. Pourtant, la couverture lexicale obtenue avec ce vocabulaire est moindre.

8.2.6 Expérimentation EuroNews

Les deux expérimentations précédentes ont montré la sensibilité des méthodes proposées dans cette thèse, à la pertinence des données d'adaptation. Cette troisième série d'expériences est effectuée sur des émissions du mois de novembre 2003. Elle a pour objectif de se placer dans le cas où les données d'adaptation sont pertinentes, puisqu'elles sont constituées à partir de transcriptions approximatives d'émissions similaires qui précèdent celle que l'on souhaite transcrire.

Description des corpus

EuroNews est une chaîne d'information qui couvre l'ensemble de l'actualité mondiale d'un point de vue européen et ce, en sept langues. La chaîne EuroNews diffuse chaque jour trois émissions d'actualité. Pour chaque sujet constituant les émissions, la vidéo est disponible sur le site Web de la chaîne[7] ainsi que sa transcription approximative. Le corpus de test est constitué à partir des sujets disponibles sur le site Web sur 4 jours du mois de novembre 2003. Par jour, cela correspond à environ 2 500 mots pour 15 minutes de signal. Les données d'évaluation sont construites à partir des transcriptions approximatives disponibles sur le site Web. Ces transcriptions sont nettoyées et normalisées selon la méthode utilisée pour les textes extraits du Web décrite au paragraphe 6.1.1.

Un premier corpus d'entraînement est $Adapt_{28}(j)$ décrit au paragraphe 8.2.1, qui est constitué à partir des textes de presses extraits du Web. A cela s'ajoute le corpus $EuroNews_{28}(j)$ qui contient les textes téléchargés à partir du site de la chaîne EuroNews, datant du jour même (antérieur strictement à l'émission traitée) et des 27 jours précédents. Le corpus de développement utilisé par l'algorithme vectoriel d'adaptation du vocabulaire, est constitué à partir des données du site EuroNews considérées sur une semaine ($EuroNews_7(j)$).

Évaluation

Les transcriptions approximatives disponibles sur le site Web sont utilisées pour évaluer les résultats de reconnaissance. Les différences entre ces transcriptions et le contenu réel des documents audiovisuels se quantifie principalement en terme de phrase. La transcription automatique contient une ou plusieurs phrases en plus, ou en moins. La procédure d'alignement effectuée par les outils d'évaluations du NIST permet de résoudre en partie le problème en comparant les phrases qui se correspondent entre elles. Par contre les phrases apparaissant en plus ou en moins, augmentent de manière artificielle le taux d'insertion ou le taux d'élision. C'est pourquoi, la mesure retenue pour évaluer les résultats de reconnaissance n'est pas le taux d'erreur, mais le taux de substitution. Cependant, un sous-corpus de test, constitué des émissions d'une journée, a été transcrit manuellement afin de pouvoir estimer le WER réel, et ainsi vérifier la corrélation entre le taux de substitution estimé sur les transcriptions approximatives, et le WER.

Expérimentation

L'adaptation du modèle de langage est effectuée via l'interpolation linéaire décrite au paragraphe 8.2.2. Dans un premier temps le vocabulaire de référence est utilisé, puis dans un second temps le vocabulaire est adapté selon la méthode vectorielle. Les résultats sont rassemblés dans le tableau 8.10.

Afin de dissocier l'influence des différentes données d'entraînement, deux adaptations sont d'abord expérimentées qui utilisent le vocabulaire de référence. La première utilise comme corpus d'adaptation pour le modèle $M_a(j)$ le corpus $Adapt_{28}(j)$ constitué uniquement de textes de presse. Le gain en terme de taux de substitution est de 1,2% absolu. Lorsque les données $EuroNews_{28}(j)$ sont ajoutées au corpus d'entraînement de $M_a(j)$, le gain est nettement plus important puisqu'il est de 3,4% absolu par rapport aux résultats obtenues avec le modèle de référence. Ainsi l'utilisation des données provenant du site d'EuroNews est déterminante.

[7] http ://www.euronews.fr

Vocabulaire	*Corpus d'adaptation*	*%MHV*	*%Sub*
Référence	Aucun	2,2	16,9
	$A_{28}(j)$	2,2	15,7
	$A_{28}(j)$ et EuroNews	2,2	13,5
Vectoriel	$A_{28}(j)$ et EuroNews	1,1	12,6

TAB. 8.10 – Résultats de reconnaissance obtenus sur le corpus EuroNews avec le vocabulaire et le modèle de référence, puis différentes configurations d'adaptation du modèle de langage.

	Référence	*Adapté*
%Substitution	18,5	13,0
%WER	29,9	23,0

TAB. 8.11 – Résultats de reconnaissance obtenus sur le sous-corpus de test EuroNews avec le système de référence et le système adapté (vocabulaire et modèle de langage). Les résultats sont exprimés en terme de taux de substitution et de WER.

L'adaptation vectorielle du vocabulaire montre, dans le cadre de ces expérimentations, son efficacité puisque le taux de MHV est divisé par deux. L'adaptation du modèle de langage qui utilise le vocabulaire adapté et comme corpus d'adaptation $[Adapt_{28}(j) \cup EuroNews_{28}(j)]$ permet d'obtenir un gain relatif de 25% sur le taux de substitution.

L'usage du sous-corpus de test permet de comparer le gain sur le taux de substitution à l'aune du WER. Comme le montre le tableau 8.11, les gains absolus sont équivalents (5,5% sur le taux de substitution et 6,9% sur le taux d'erreur). Ainsi l'observation du taux de substitution sur les transcriptions approximatives permettent une estimation des gains potentiels sur le WER. Cette série d'expérimentations confirme donc l'impact du choix des données d'adaptation sur les performances des méthodes proposées pour l'adaptation du vocabulaire et du modèle de langage. Les résultats obtenus valident l'approche dans le cas où des données d'adaptations pertinentes sont disponibles, et montrent à nouveau l'acuité de l'algorithme d'adaptation vectoriel du vocabulaire.

Conclusion

L'adaptation du modèle de langage a été abordée selon deux problématiques spécifiques à l'indexation audiovisuelle : le traitement des archives et du flux entrant. Dans les deux cas, le Web a été utilisé pour constituer les ressources textuelles nécessaires à l'adaptation des modèles de langage.

La spécificité du traitement des archives se situe autant sur un niveau linguistique qu'acoustique. Les expérimentations sur la modélisation linguistique montrent l'impact du décalage entre l'époque du corpus d'entraînement du modèle de référence (principalement les années 90) et l'époque du corpus de test (des années 40 à 90). Lorsque ce décalage est de 40 ans, il se traduit par une dégradation de près de 40% de la perplexité calculée avec le modèle de référence. La méthode d'adaptation proposée effectue l'interpolation linéaire du modèle de référence avec le modèle d'époque estimé à partir de textes extraits du Web. Le coefficient d'interpolation est calculé par décennie de manière non-supervisée, en minimisant la perplexité de transcriptions automatiques préalablement générées avec le modèle de référence. Les gains en perplexité varient suivant les décennies. Il est de 9% avec les données des années 40, et il décroît jusqu'à s'annuler avec les données des années 90. L'adaptation des modèles de langage permet un gain significatif du taux d'erreur uniquement pour les années 40, et sur l'ensemble du corpus de test, l'adaptation n'a pas d'impact sur le taux d'erreur. Les autres travaux menés sur le corpus ECHO et relatés dans l'article [Barras *et al.*, 2002] montrent que les conditions acoustiques constituent la première source de difficultés et d'erreurs pour un système de reconnaissance automatique de la parole .

Pour le traitement du flux entrant, un modèle de langage tournant est mis en place pour transcrire automatiquement des émissions d'actualité au jour le jour. L'objectif est de pouvoir adapter le système aux évolutions thématiques de l'actualité qui se traduisent par des changements lexicaux et linguistiques. Au chapitre 6, plusieurs algorithmes d'adaptation du vocabulaire ont été présentés. Dans ce chapitre deux d'entre eux, les algorithmes *ad hoc* et vectoriels, sont retenus pour adapter le vocabulaire. L'adaptation du modèle de langage remplit donc un double objectif : d'abord ajouter les mots nouveaux dans le modèle en leur attribuant les probabilités nécessaires, puis d'introduire des informations linguistiques liées l'actualité.

L'adaptation effectue l'interpolation du modèle de référence avec un modèle spécifique au jour. L'interpolation linéaire et l'interpolation sélective sont les deux méthodes proposées qui permettent de maintenir le vocabulaire du système sous la limite des 65 535 mots. Deux corpus de test distincts sont utilisés qui permettent d'explorer pour RFI, le cas où des données d'adaptation proviennent du site du diffuseur, et pour TF1, le cas où il n'y a pas de données spécifiques à la tache.

Les résultats montrent que l'interpolation sélective permet d'obtenir des meilleurs résultats dans tous les cas de figure. L'impact du choix des corpus de développement et d'entraînement peut être observé sur la couverture lexicale, la perplexité et le taux d'erreur de reconnaissance. Sur le corpus de test composé d'émissions diffusées par RFI, les gains en terme de taux d'erreur sont en moyenne de 0,8% (soit de 4,0% relatif), alors que sur le corpus composé d'émissions diffusées par TF1, il n'est plus que de 0,5% (soit 2,1% relatif). Les gains relatifs peuvent donc être divisés par deux suivant le corpus de test choisi. Les méthodes proposées dépendent donc fortement du choix des corpus de développement et de leurs représentativités de la tâche. L'expérimentation sur les données EuroNews de novembre 2003 confirme cela. L'utilisation, pour les données d'adaptation, de transcriptions approximatives d'émissions similaire provenant du site Web du diffuseur, permet d'obtenir un gain de 50% sur la couverture lexicale et de 25% sur le taux de substitution.

Conclusion

L'index est au nombre de ces usages qui restent comme des témoins des anciens temps au milieu des temps nouveaux.[8]

Les systèmes d'indexation automatique ont pour objectif d'extraire des informations pertinentes permettant de retrouver ultérieurement un document audiovisuel dans un ensemble. Les travaux effectués au cours de cette thèse s'inscrivent dans la volonté de faciliter l'extraction d'informations lexicales et linguistiques de la bande sonore d'un document audiovisuel en vue de son indexation. Pour fonder la pertinence de ces informations, je me suis attaché à développer des méthodes d'adaptation du vocabulaire et des modèles de langage d'un système de transcription automatique de la parole.

Dans le premier chapitre, l'indexation audiovisuelle est présentée dans le cadre particulier de l'INA, où cette pratique documentaire est confrontée à la variété des émissions, des niveaux de description requis, et à la quantité de données à traiter. Dans ce contexte, les techniques d'indexation automatique de documents audiovisuels, telles qu'elles sont présentées au deuxième chapitre, ne semblent pas encore suffisamment robustes pour remplacer, en l'état, l'indexation manuelle. Les campagnes d'évaluation menées par le NIST ont fédéré les recherches en indexation automatique. Plus particulièrement, la campagne TRECVID est dédiée à l'indexation automatique et à la recherche par le contenu de documents audiovisuels. En comparant des systèmes faisant intervenir différentes techniques d'analyse automatique de l'image (par exemple la détection de visages ou de changement de plan) et de l'audio (la transcription automatique ou la segmentation parole/musique/bruit), cette campagne établit que l'état de l'art en matière d'indexation automatique de documents audiovisuels est constitué d'un système de type *SDR* (*Spoken Document Retrieval*) qui allie un système de reconnaissance automatique de la parole à un moteur de recherche.

Ainsi, la transcription automatique de la parole est l'accès premier au contenu audiovisuel et les erreurs, qu'elle peut contenir, en déterminent la pertinence. Afin d'avoir une vue d'ensemble du processus de reconnaissance de la parole et de situer les sources d'erreurs, le système de transcription automatique du LIMSI et ses différentes composantes ont été présentés au chapitre 3. Cette thèse s'inspire des conclusions du NIST sur les systèmes *SDR* qui soulignent la nécessité d'adapter le vocabulaire et le modèle de langage à l'émission à transcrire. Au cours du chapitre 4,

[8]Chateaubriand, Mémoires d'Outre tombe.

les principes théoriques de la modélisation statistique du langage ont été introduits, et les aspects pratiques de la construction d'un modèle de langage y ont été ensuite développés. Enfin les techniques d'adaptation existantes pour la transcription automatique de documents audiovisuels ont été présentées. Puisque la disponibilité de données d'adaptation est souvent problématique, les techniques d'adaptation sont regroupées selon les données d'adaptation qu'elles requièrent. Ces deux chapitres 3 et 4 permettent de définir les cadres théoriques et pratiques utilisés dans la suite de ces travaux.

Le vocabulaire est la liste finie (limitée généralement à 65 535 mots) des mots connus par le système à partir de laquelle sont construits le dictionnaire de prononciation et le modèle de langage. Dans une perspective d'indexation, son rôle détermine le contenu possible des transcriptions automatiques. Dans le chapitre 5, les cadres théoriques et d'évaluation spécifiques à la construction du vocabulaire sont présentés. Ils permettent de définir l'ensemble des contraintes à respecter, et le *modus operandi*. La construction manuelle du vocabulaire de référence est détaillée, et une typologie des mots hors vocabulaire est effectuée établissant que plus de la moitié d'entre eux sont des noms propres. Ensuite, est proposé un algorithme de construction automatique qui interpole linéairement des distributions unigrammes estimées sur des données d'entraînement, de manière à minimiser la perplexité d'un corpus de développement (transcriptions manuelles de journaux télévisés de 2000). Les données utilisées sont exactement les mêmes que celles qui ont servi pour le vocabulaire de référence. Plusieurs segmentations des données d'entraînement sont expérimentées. Dans le meilleur des cas, l'algorithme automatique construit un vocabulaire conduisant à une diminution relative du taux de mots hors vocabulaire de 17% sur le corpus de développement. Les résultats observés sur un corpus de test de deux ans plus récent (janvier 2002) illustrent le besoin d'utiliser des données d'entraînement contemporaines de l'émission à transcrire.

Étant donné le coût important lié à la préparation des corpus d'apprentissage, le vocabulaire et les modèles de langage utilisés par un système de transcription sont généralement développés à partir d'un corpus figé, qui ne peut qu'exceptionnellement être contemporain du document à transcrire. Que ce soit pour le vocabulaire ou le modèle de langage de référence, les données d'entraînement les plus récentes datent de l'an 2000. Pour transcrire des journaux télévisés contemporains, le décalage temporel est de trois ans. Or les techniques de construction présentées aux chapitres 4 et 5 ne sont pas appropriées à une mise à jour des modèles linguistiques, puisqu'elles nécessitent des données d'apprentissage en quantité souvent inaccessible.

Au chapitre 6, sont exposées trois méthodes d'adaptation du vocabulaire d'un système de transcription, à partir de textes disponibles sur internet. Ces textes sont collectés quotidiennement à partir de plusieurs sites, et une méthode spécifique transformant ces fichiers HTML en un corpus utilisable est appliquée. Pour chaque jour du mois de janvier 2002, un corpus d'adaptation, un corpus de développement et un corpus de test sont constitués afin de mener une expérimentation et une évaluation quotidienne. Les deux méthodes qui se distinguent sont celle appelées *ad hoc* et vectorielle. Elles permettent une diminution relative du taux de mots hors vocabulaire respectivement de 32% et de 37 %.

L'étude des performances obtenues avec les différents vocabulaires pour chaque jour du mois de janvier montre que les mots de l'actualité changent suffisamment rapidement pour justifier une adaptation quotidienne. De plus, le vocabulaire construit avec l'algorithme vectoriel est localement le plus performant mais est dépendant du corpus de développement, alors que le vocabulaire construit avec l'algorithme *ad hoc* conduit à une adaptation moins performante mais plus pérenne. Plutôt que de travailler avec un vocabulaire de taille fixe, la couverture lexicale peut être fixée pour dimensionner sa taille. Les expériences montrent que l'adaptation vectorielle implique en moyenne une division de la taille du vocabulaire par deux pour une couverture lexicale de 98,5 %.

Tout changement dans le vocabulaire entraîne la modification du modèle de langage afin d'attribuer des probabilités aux mots nouveaux. Dans le chapitre 7, la technique du repli lexical est introduite, pour construire un modèle de langage dont le vocabulaire est ouvert. L'objectif est d'autoriser l'introduction de mots nouveaux sans réentraîner le modèle. Deux cadres d'expérimentation sont proposés pour évaluer cette méthode. Dans le premier, les mots nouveaux sont sélectionnés manuellement et introduits dans le vocabulaire de manière à supprimer la présence de mots hors vocabulaire. Cette expérimentation d'oracle permet d'évaluer les capacités de la méthode à transcrire les mots ajoutés. Le taux moyen de reconnaissance sur les mots introduits via le repli lexical est de 78,7%. La réduction relative du WER est de 10%, montrant ainsi un rapport moyen de 2,1 entre le gain en couverture lexicale et la réduction au niveau du WER. Le second cadre d'application consiste à utiliser le repli lexical pour introduire les mots nouveaux qui sont sélectionnés automatiquement par la méthode *ad hoc* d'adaptation du vocabulaire. Les résultats de reconnaissance obtenus lors de cette expérimentation montrent une dégradation des performances du système de reconnaissance, et mettent en évidence les limitations du repli lexical dues principalement à l'estimation des probabilités des formes de repli.

Dans le chapitre 8 sont proposées des méthodes d'adaptation des modèles de langage appliquées d'une part à un corpus d'archive, et d'autre part, à la transcription quotidienne d'émissions d'actualité. Dans les deux cas, les données d'adaptation sont extraites de sites internet. L'analyse des erreurs de transcription automatique obtenues avec le système de référence sur le corpus ECHO permet de définir les problématiques (de nature acoustique et linguistique) spécifiques au traitement des archives. La méthode d'adaptation des modèles de langage qui est ensuite proposée effectue l'interpolation linéaire du modèle de référence avec un modèle d'époque estimé à partir de textes extrait du Web. Le coefficient d'interpolation est calculé par décennie de manière non-supervisée, en minimisant la perplexité de transcriptions automatiques préalables. La réduction relative de la perplexité est de 9% pour les données des années 40 et décroît jusqu'à s'annuler pour les années 90 pour lesquelles le modèle de référence est le plus approprié.

Pour la transcription quotidienne d'émissions d'actualité, un modèle de langage tournant est mis en place. L'objectif est de pouvoir adapter le système aux évolutions thématiques de l'actualité qui se traduisent par des changements lexicaux et linguistiques. L'expérimentation porte sur le mois de janvier 2002. Au chapitre 6, plusieurs algorithmes d'adaptation du vocabulaire ont été présentés. Dans ce chapitre, deux d'entre eux sont retenus pour adapter le vocabulaire. L'adaptation du modèle de langage remplit donc un double objectif : d'une part ajouter les mots nouveaux dans le modèle en leur attribuant les probabilités nécessaires, et d'autre part, introduire des informations linguistiques caractéristiques de l'actualité. L'adaptation effectue l'interpolation du modèle de référence avec un modèle propre à chaque jour. Deux méthodes d'interpolation sont proposées qui permettent de maintenir le vocabulaire du système sous la limite des 65 535 mots :

l'interpolation linéaire, et l'interpolation sélective.

Dans la première série d'expériences, deux corpus de test distincts sont constitués, l'un à partir d'émissions diffusées par RFI, et l'autre à partir de journaux télévisés de TF1. Ces deux corpus permettent d'explorer pour RFI, le cas où des données d'adaptation provenant du site du diffuseur sont disponibles, et pour TF1, le cas contraire. Sur le corpus de test composé d'émissions diffusées par RFI, les gains en terme de taux d'erreur sont en moyenne de 0,8% (soit de 4,0% relatif), alors que sur le corpus composé d'émissions diffusées par TF1, il n'est plus que de 0,5% (soit 2,1% relatif). Les gains relatifs peuvent donc être divisés par deux suivant le corpus de test utilisé. Les méthodes d'adaptation proposées montrent ainsi leur sensibilité à la pertinence des données d'adaptation. Dans la seconde série d'expériences, pour transcrire des émissions d'actualité diffusées par EuroNews, des transcriptions approximatives d'émissions similaires diffusées par la même chaîne, sont utilisées pour adapter le vocabulaire et le modèle de langage. Cela permet d'envisager le cas où les données d'adaptation sont effectivement pertinentes. Un gain important de 25% sur le taux de substitution est observé en utilisant les transcriptions approximatives pour l'évaluation.

Ainsi, d'un point de vue pratique, la problématique abordée dans cette thèse est l'introduction d'informations dans la modélisation linguistique. Implicitement les choix de modélisation ne sont pas remis en cause (modèle *ngram* de mots). Mais les deux problèmes soulevés sont quelles informations est-il utile d'introduire et comment.

L'information (ou les connaissances) introduite est obtenue à partir d'observations effectuées sur les corpus de développement et d'entraînement. Leurs constitution est donc déterminante. En effet que ce soit pour l'adaptation du vocabulaire ou du modèle de langage, les méthodes développées dépendent du choix de ces corpus. Ce choix se fonde sur des hypothèses de corrélation de contenus entre les textes extraits du Web et les émissions à transcrire. Les textes sont alors considérés comme constituant une connaissance *a priori*. Les résultats des chapitres 6 et 8 montrent à la fois la fragilité de ces hypothèses et leur impact sur les performances. Pour une application réaliste des techniques proposées, il est donc opportun d'orienter les recherches vers des méthodes d'extraction à partir d'un document audio, d'informations permettant de choisir un matériau d'adaptation approprié pour l'application cible. L'objectif est alors de s'affranchir de la nécessité d'une connaissance *a priori* sur le document à transcrire qui est loin d'être toujours disponible.

Pour sélectionner les données d'adaptation, il est difficile d'extraire des informations fiables à partir des résultats du système de reconnaissance alors qu'ils contiennent des erreurs. Deux solutions sont envisageables (les deux peuvent être complémentaires) : l'une consiste à sélectionner les informations en fonction d'une mesure de confiance ; l'autre tente de généraliser les connaissances contenues dans les résultats produits par le système de reconnaissance et ainsi combler les lacunes créées par les erreurs. En ce qui concerne l'estimation d'une mesure de confiance, le treillis généré par le système de reconnaissance est un point de départ prometteur puisqu'il représente l'espace de recherche du système de reconnaissance, et qu'il contient toutes les informations nécessaires au décodage (acoustiques et linguistiques). La problématique de la généralisation à partir des résultats de reconnaissance, a déjà été brièvement introduit au paragraphe 4.5.2. Les recherches peuvent alors s'inspirer des techniques de recherches et d'extraction d'information et en particulier de l'analyse *LSA* (*Latent Semantic Analysis*) telle qu'elle est succinctement présentée au paragraphe 4.5.3. Dans ce cas, le treillis semble également approprié puisqu'il permet d'envisager des hypothèses concurrentes ce qui augmentent la probabilité d'inclure l'hypothèse juste.

Une fois les données d'adaptation sélectionnées, la question de l'adaptation du modèle de langage se pose. Les résultats du chapitre 8 montrent les faiblesses des méthodes d'adaptation des modèles de langage proposées. Lors de l'adaptation du modèle de langage de référence, un des inconvénients de l'interpolation linéaire est qu'elle modifie l'ensemble du modèle de référence. Cette modification ne tient pas compte de la nature des ngrammes (sont-ils spécifiques ou appartiennent-ils au français général), et de la pertinence de l'estimation de leurs probabilités. L'interpolation sélective proposée au chapitre 8 tente donc de sélectionner les ngrammes dont les probabilités doivent être modifiées. Mais le gain apporté par cette méthode est timide et incite à développer une adaptation faisant preuve d'une sélectivité plus marquée avec par exemple une interpolation où le coefficient est dépendant du contexte.

La notion de sélectivité renvoie à un problème bien plus fondamental qui est l'absence de structure d'un modèle de langage classique. Ce problème se pose entre autre au stade de l'adaptation. L'objectif est alors d'être à même de généraliser des connaissances acquises grâce à des observations lacunaires dues à l'utilisation de corpus souvent de taille réduite. Or pour un modèle *ngram* de mots, l'espace de représentation est l'espace discret des indices des mots. Cet espace est construit sans aucune considération sémantique ou linguistique, et les indices n'entretiennent aucunes relations entre eux. Les modèles connexionistes décrits au paragraphe 4.1.3 proposent de se placer dans un espace continu pour estimer la probabilité d'un mot en contexte. L'utilisation de cet espace introduit la notion de proximité entre mots et autorise donc, dans une perspective d'adaptation, la possibilité d'effectuer des transformations dans cet espace qui tiennent compte de la distance entre les mots. D'autres solutions sont envisageables pour structurer les modèles de langage. Si l'utilisation des catégories morpho-syntaxiques (évoquée au paragraphe 4.1.2) ne semble pas spécifiquement appropriée à la transcription automatique de documents audiovisuels, d'autres types d'analyses peuvent être envisagée comme l'analyse par contrainte (utilisation des *Constraint Dependency Grammar*) [Wang and Harper, 2002].

L'extraction d'informations à partir d'un document audio permettant de rassembler des ressources textuelles liées au document offre d'autres opportunités que la constitution d'un corpus d'adaptation. Les textes ou articles ainsi sélectionnés peuvent enrichir l'indexation de ce document, en proposant un matériau pour la construction de métadonnées. En particulier, il semble intéressant d'être à même d'aligner ou d'établir des relations de dépendances entre ces textes et le contenu du document à transcrire. La perspective est alors de structurer, non seulement le document audio, mais également des connaissances complémentaires sur son contenu.

Bibliographie

[Abberley *et al.*, 1999] D. Abberley, S. Renals, Dan Ellis, and T. Robinson. The THISL SDR system at trec-8. In *Proceedings of the 8th Text Retrieval Conference TREC-8*, Gaithersburg, MD, November 1999.

[Adda and Adda-Decker, 1997] G. Adda and M. Adda-Decker. Normalisation de textes en français : une étude quantitative pour la reconnaissance de la parole. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 289–296, Avignon, France, 1997.

[Adda *et al.*, 1997a] G. Adda, M. Adda-Decker, J. L. Gauvain, and L. Lamel. Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 35–40, Avignon, France, April 1997.

[Adda *et al.*, 1997b] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel. Text normalization and speech recognition in French. In *Proc. Eurospeech*, volume 5, pages 2711–2714, Rhodes, September 1997.

[Adda-Decker *et al.*, 1999] Martine Adda-Decker, Gilles Adda, Jean-Luc Gauvain, and Lori Lamel. Large vocabulary speech recognition in french. In *Proc. ICASSP*, Phoenix, March 1999.

[AFNOR, 1996] AFNOR. *Information et documentation, principes généraux pour l'indexation des documents*, volume Documentation, présentation des publications, traitements documentaire et gestion de bibliothèques. Association Française de Normalisation, 1996.

[Allauzen and Gauvain, 2002] Alexandre Allauzen and Jean-Luc Gauvain. Mise à jour automatique du modèle de langage d'un système de transcription. In *Actes des XXIIièmes Journées d'Études sur la Parole*, pages 305–308, Nancy, France, June 2002.

[Allauzen and Gauvain, 2003] Alexandre Allauzen and Jean-Luc Gauvain. Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés. *Traitement Automatique des langues*, 44(1) : 11–31, 2003.

[Auffret, 2000] Gwendal Auffret. *Structuration de documents audiovisuels et publication électronique.* PhD thesis, Université Technologique de Compiègne, Institut National de l'Audiovisuel, 2000.

[Auzanne *et al.*, 2000] Cedric Auzanne, John S. Garofolo, Jonathan G. Fiscus, and William M. Fisher. Automatic language model adaptation for spoken document retrieval. In *SDR 2000.* TREC 9, 2000.

[Bachimont, 98] Bruno Bachimont. Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. *Document numérique*, 2(3-4) :219–242, 98.

[Bahl *et al.*, 1977] Lalit Bahl, Jim Baker, Frederick Jelinek, and Robert Mercer. Perplexity : a measure of the difficulty of speech recognition task. In *Program of the 94th Meeting of the Acoustical Society of America*, volume 62 :S63, Miami Beach, Florida, december 1977.

[Barras *et al.*, 2001] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2) :5–22, 2001.

[Barras *et al.*, 2002] C. Barras, A. Allauzen, L. Lamel, and Jean-Luc Gauvain. Transcribing audio-video archives. In *Proc. ICASSP*, volume 1, pages 13–16, Orlando, Florida, May 2002.

[Baum *et al.*, 1970] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41 : 164–171, 1970.

[Bazzi, 2002] Issam Bazzi. *Modeling Out-Of-Vocabulary Words For Robust Speech Recognition*. PhD thesis, Massachusetts Institute of Technology, 2002.

[Béchet and Yvon, 2000] Frédéric Béchet and François Yvon. Les noms propres en traitement automatique de la parole. *Traitement automatique des langues*, 41(3) : 671–709, 2000.

[Beaujard, 2000] Christel Beaujard. *Modèles de langage fondés sur des classes et des concaténations de mots pour la reconnaissance de la parole dans des systèmes de dialogue Homme/Machine*. PhD thesis, Université Paris-Sud Orsay, 2000.

[Beeferman *et al.*, 1997] Doug Beeferman, Adam Berger, and John Lafferty. A model of lexical attraction and repulsion. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 373–380, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[Bengio and Ducharme, 2001] Yoshua Bengio and Réjean Ducharme. A neural probabilistic language model. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13. Morgan Kaufmann, 2001.

[Berger *et al.*, 1996] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1) : 39–71, 1996.

[Bolle *et al.*, 1998] R. M. Bolle, B.-L. Yeo, and M. M. Yeung. Video query : Research directions. *IBM Journal of Research and Development*, 42(2) : 233–252, 1998.

[Brill *et al.*, 1998] E. Brill, R. Florian, J. C. Henderson, and L. Mangu. Beyond n-grams : Can linguistic sophistication improve language modeling ? In *Proceedings of COLING-ACL 98*, Quebec, Canada, August 1998.

[Brill, 1994] Eric Brill. Some advances in rule based part-of-speech tagging. In AAAI, editor, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722–727, Seattle, WA, 1994.

[Brown *et al.*, 1990] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2) :79–85, 1990.

[Brown *et al.*, 1992] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4) :467–479, 1992.

[Chen and Goodman, 1996] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In Arivind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[Chen and Goodman, 1998] Stanley F. Chen and Joshua T. Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[Chen *et al.*, 1998] Stanley F. Chen, Douglas Beeferman, and Ronald Rosenfeld. Evaluation metrics for language models. In *In DARPA Broadcast News Transcription and Understanding Workshop.*, 1998.

[Chen *et al.*, 2001] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker. Language Model Adaptation for Broadcast News Transcription. In *Proc. ISCA ITRW 2001 Adaptation Methods for Speech Recognition*, Sophia-Antipolis, August 2001.

[Clarkson and Robinson, 1997] P. Clarkson and A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *Proc. ICASSP '97*, pages 799–802, Munich, Germany, 1997.

[Clarkson and Robinson, 1998] Philip Clarkson and Tony Robinson. The applicability of adaptive language modelling for the broadcast news task. In *ICSLP*, pages 962–965, Sydney, Australia, 1998.

[Clarkson and Robinson, 1999] Philip Clarkson and Tony Robinson. Towards improved language model evaluation measures. In *Eurospeech 99*, 1999.

[Clarkson and Robinson, 2001] Philip Clarkson and Tony Robinson. Improved language modelling through better language model evaluation measures. *Computer Speech and Language*, 15(1) :39–53, 2001.

[Clarkson and Rosenfeld, 1997] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-cambridge toolkit. In *Proc. Eurospeech '97*, pages 2707–2710, Rhodes, Greece, 1997.

[Cover and Thomas, 1991] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* New York, USA, 1991.

[Davis and Mermelstein, 1980] S. Davis and P Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech and Audio Processing*, 28(4) :357–366, 1980.

[de Jong *et al.*, 1999] F. de Jong, J.L. Gauvain, J. den Hartog, and K. Netter. Olive : Speech based video retrieval. In *Proc. CBMI'99*, Toulouse, Oct 1999.

[de Mareüil, 1996] P. Boula de Mareüil. Pour une approche par règles en transcription graphème-phonème. In *Séminaire GDR-PRC CHM Lexique et communication parlée*, Toulouse, 1996.

[de Mori and Bigi, 2002] R. de Mori and Brigitte Bigi. Principe de reconnaissance. In J. Mariani, editor, *Reconnaissance de la parole Traitement automatique du langage parlé*, volume 2, pages 21–46. Hermes Lavoisier, 2002.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1) :1–38, november 1977.

[Duda and Hart, 1974] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1974.

[Federico and Bertoldi, 2001] Marcello Federico and Nicola Bertoldi. Broadcast news adaptation using contemporary texts. In *Proc. Eurospeech*, pages 239–242, Aalborg, Denmark, September 2001.

[Federico, 1996] M. Federico. Bayesian estimation methods for N-gram language model adaptation. In *Proc. ICSLP '96*, volume 1, pages 240–243, Philadelphia, PA, 1996.

[Federico, 1999] M. Federico. Efficient language model adaptation through mdi estimation. In *Proc. of EUROSPEECH*, pages 1583–1586, Budapest, Hungary., 1999.

[Federico, 2002] Marcello Federico. Language model adaptation through topic decomposition and mdi estimation. In *Proc. ICASSP*, volume 1, pages 773–776, Orlando, Florida, May 2002.

[Garofolo *et al.*, 1998] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, Cedric G. P. Auzanne, and Bruce Z. Lund. 1998 spoken TREC-7 document retrieval track overview and results. In *Text REtrieval Conference*, pages 79–90, 1998.

[Garofolo *et al.*, 1999] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track : A success story. In *Proceedings of the 8th Text Retrieval Conference TREC-8*, pages 107–130, Gaithersburg, Maryland, November 1999.

[Gauvain and Lamel, 2002] J.L. Gauvain and L.F. Lamel. Systèmes de reconnaissance, de compréhension et de dialogue. In J. Mariani, editor, *Reconnaissance de la parole Traitement automatique du langage parlé*, volume 2, pages 47–83. Hermes Lavoisier, 2002.

[Gauvain and Lamel, 2003] Jean-Luc Gauvain and Lori Lamel. Large vocabulary speech recognition based on statistical methods. In W. Chou and F. Juang, editors, *Pattern Recognition in Speech and Language Processing*, chapter 5, pages 149–189. CRC Press, 2003.

[Gauvain and Lee, 1994] J. Gauvain and C. Lee. Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2, 1994.

[Gauvain *et al.*, 2000] J.L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. Kercadio. The LIMSI SDR system for TREC-9. In *Proc. of the Text Retrieval Conference, TREC-9*, pages 335–341, Gaithersburg, Nov 2000.

[Gauvain *et al.*, 2001] J.-L. Gauvain, L. Lamel, and G. Adda. Audio partitioning and transcription for broadcast data indexation. *MTAP Journal*, 14(2) :187–200, 2001.

[Gauvain *et al.*, 2002] Jean-Luc Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37 :98–108, 2002.

[Gauvain *et al.*, 2003] J.L. Gauvain, L. Lamel, G. Adda, L. Chen, and H. Schwenk. The limsi rt03 bn systems. In *Proceedings of Riched Transcription 2003 Summer Workshop*, Boston, MA, USA, May 2003.

[Good, 1953] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3, 4) :237–264, 1953.

[Goodman, 1976] Robert Gary Goodman. Analysis of languages for man-machine voice communication. Technical report, Carnegie Mellon University, School of Computer Science, 1976. PhD Thesis.

[H. Ney, 1994] R. Kneser H. Ney, U. Essen. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1) :1–38, 1994.

[Hermansky, 1990] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4) :1738–1752, 1990.

[Hunter and Armstrong, 1999] Jane Hunter and Liz Armstrong. A comparison of schemas for video metadata representation. *Computer Networks (Amsterdam, Netherlands : 1999)*, 31(11–16) :1431–1451, May 1999.

[Iyer *et al.*, 1997] R. Iyer, M. Ostendorf, and M. Meteer. Analysing and predicting language model improvements. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Undestanding*, pages 254–261, 1997.

[Jelinek and Mercer, 1980] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands, May 1980.

[Jelinek, 1976] Frederick Jelinek. Continuous speech recognition by statistical methods. *Proceedings of IEEE*, 64(4) :532–556, April 1976.

[Jelinek, 1990] F. Jelinek. *Self-Organized Language Modeling for Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers Inc., 1990.

[Jelinek, 2000] Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 2000.

[Johnson *et al.*, 2001] S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland. Spoken Document Retrieval for TREC-9 at Cambridge University. In *Proc. 'The Ninth Text REtrieval Conference (TREC-9)'.*, Gaithersburg, MD, 2001.

[Katz., 1987] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 35(3) :400–401, march 1987.

[Kemp and Waibel, 1998] T. Kemp and A. Waibel. Reducing the oov rate in broadcast news speech recognition. In *Proc. ICSLP*, volume 5, pages 1839–1842, Sydney, Australia, 1998.

[Kim *et al.*, 2003] Do Yeong Kim, Gunnar Evermann, Thomas Hain, David Mrva, Sue Tranter, Lan Wang, Phil Woodland, , and Rest of the HTK STT team. 2003 cu-htk broadcast news english system development. In *Proceedings of Riched Transcription 2003 Summer Workshop*, Boston, MA, USA, May 2003.

[Kneser and Ney, 1993] R. Kneser and H. Ney. Improved clustering techniques for class-based statistical language modelling. In *In Proc. European Conf. on Speech Technology*, pages 973–976, 1993.

[Kneser and Ney, 1995] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram lan- guage modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184, 1995.

[Kneser *et al.*, 1997] R. Kneser, J. Peters, and D. Klakow. Language model adaptation using dynamic marginals. In *Proceedings of Eurospeech'97*, pages 1971–1974, 1997.

[Kuhn and Mori, 1990] R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6) :570–583, june 1990.

[Lee and Seung, 2000] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[Lefevre *et al.*, 2001] Fabrice Lefevre, Jean-Luc Gauvain, and Lori Lamel. Towards Task-Independent Speech Recognition. In *Proc. ICASSP*, volume 1, page 521, Salt Lake City, May 2001.

[Leggetter and Woodland, 1995] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2), 1995.

[Mangu *et al.*, 2003] L. Mangu, G. Zweig, and B. Ramabhadran. Sausage comprehension and listening experiments. In *Proceedings of Riched Transcription 2003 Summer Workshop*, Boston, MA, USA, May 2003.

[Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[Martin *et al.*, 1995] S. Martin, J. Liermann, and H. Ney. Algorithms for bigram and trigram word clustering. In *Proceedings of Eurospeech-95*, pages 1253–1256, 1995.

[Nguyen *et al.*, 2003] Long Nguyen, Nicolae Duta, John Makhoul, Spyros Matsoukas, Richard Schwartz, Bing Xiang, and Dongxin Xu. The bbn rt03 bn english system. In *Proceedings of Riched Transcription 2003 Summer Workshop*, Boston, MA, USA, May 2003.

[Niesler *et al.*, 1998] T. Niesler, E. Whittaker, and P. Woodland. Comparison of part-of-speech and automatically derived category-based language models for speech recognition. In *icassp 98*, volume 1, Seattle, 1998.

[Pietra *et al.*, 1992] Della Pietra, S., Della Pietra, Mercer V., R. L., and S. Roukos. Adaptive language modeling using minimum discriminant estimation. In *Proceedings of ICASSP-92*, volume 1, pages 633–636, San Francisco, March 1992.

[Ponte and Croft, 1998] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[Prouts, 1980] Bernard Prouts. Contribution à la synthèse de la parole à partir de texte : Transcription graphème-phonème en temps réel sur microprocesseur, Thèse de l'Université Paris XI, novembre 1980.

[Quénot *et al.*, 2002] G.M. Quénot, D. Moraru, L. Besacier, and P. Mulhern. Clips at trec 11 : Experiments in video retrieval. In *Proceedings of the 11th Text Retrieval Conference TREC-11*, Gaithersburg, Maryland, November 2002.

[Robertson *et al.*, 1994] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *NIST Special Publication 500-226 : Overview of the Third Text REtrieval Conference (TREC-3)*, November 1994.

[Rosenfeld and Huang, 1992] R. Rosenfeld and X. Huang. Improvements in stochastic language modeling. In *Proceedings of the Speech and Natural Language DARPA Workshop*, pages 107–111. Morgan Kaufmann Publishers, February 1992.

[Rosenfeld, 1995] R. Rosenfeld. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proceedings of Eurospeech 95*, pages 1763–1766, 1995.

[Rosenfeld, 1996] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10, 1996.

[Rosenfeld, 2000] R. Rosenfeld. Two decades of statistical language modeling : Where do we go from here, 2000.

[Salton and McGill, 1997] G. Salton and M. J. McGill. Smart and sire experimental retrieval system. In K. Sparck-Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 381–399. Morgan-Kaufmann, 1997.

[Schwenk and Gauvain, 2002] Holger Schwenk and Jean-Luc Gauvain. Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition. In *Proc. ICASSP*, pages 765–768, Orlando, May 2002.

[Seymore and Rosenfeld, 1997] Kristie Seymore and Ronald Rosenfeld. Using story topics for language model adaptation. In *Proc. Eurospeech '97*, pages 1987–1990, Rhodes, Greece, 1997.

[Shannon, 1951] C. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30 :50–64, 1951.

[Slype, 1987] Georges Van Slype. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Éditions d'Organisation, 1987.

[Smeaton *et al.*, 2002] Alan F. Smeaton, Paul Over, and R. Taban. The TREC-2002 video track report. In *Proceedings of the 11th Text Retrieval Conference TREC-11*, Gaithersburg, Maryland, November 2002.

[Smeulders *et al.*, 2000] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349–1380, 2000.

[Stolcke, 2002] A. Stolcke. Srilm - an extensible language lodeling toolkit. In *Proc. ICSLP*, pages 901–904, Denver, Colorado, 2002.

[Troncy, 2001] Raphaël Troncy. Études du manuel d'indexation commun à tous les documentalistes. Technical report, Institut National de l'Audiovisuel, april 2001.

[Veneau, 2002] Emmanuel Veneau. *Macro-segmentation multicritère et classification de séquences par le contenu dynamique pour l'indexation vidéo.* PhD thesis, IRISA Rennes, VISTA, 2002.

[Wang and Harper, 2002] W. Wang and M. P. Harper. The superarv language model : Investigating the effectiveness of tightly integrating multiple knowledge sources. In *Proceedings of Empirical Methods in Natural Language Processing*, 2002.

[Ward, 1990] W. Ward. The cmu air travel information service : Understanding spontaneous speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, 1990.

[Witten and Bell, 1991] I.H. Witten and T.C. Bell. The zero frequency problem : Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4) :1085–1094, July 1991.

[Yvon *et al.*, 1998] F. Yvon, P. Boula de Mareüil, C. d ?Alessandro, V. Aubergé., M. Bagein, G. Bailly, F. Béchet, S. Foukia, J.-P. Goldman, E. Keller, D. O ?Shaughnessy, V. Pagel, F. Sannier, J. Véronis, and B. Zellner. Objective evaluation of grapheme-to-phoneme conversion for text-to-speech synthesis in french. *Computer Speech and Language*, 12(3), 1998.

Annexes

	Vocabulaire		
Jour	8	15	29
8	0.99	0.91	0.93
15	0.94	0.88	0.87
29	0.96	0.88	0.90

Tab. 8.12 – Taux de mots hors vocabulaire obtenus avec les vocabulaires adaptés pour chaque jour via la méthode *ad hoc*, et calculés sur les transcriptions de référence

Jour	*#mots*	*Vocabulaire de référence*	*Vocabulaire adhoc*	*Vocabulaire vectoriel*
020102	2778	0.82	0.82	1.80
020103	7828	1.35	1.25	1.31
020104	2985	1.14	1.03	1.03
020105	2849	1.65	1.33	1.29
020107	2809	0.68	0.49	0.60
020108	2983	0.80	0.80	0.77
020109	2830	0.88	0.91	1.13
020110	2771	1.01	0.83	0.93
020111	2710	1.33	1.32	1.58
020114	2867	1.36	1.11	1.08
020115	8145	1.19	0.89	1.10
020116	2813	1.28	1.24	1.49
020117	2773	0.97	0.93	1.00
020118	2753	1.34	1.23	1.45
020121	2890	1.04	0.86	1.14
020122	2853	1.02	0.87	0.84
020123	2825	1.24	1.02	1.34
020124	2846	1.09	1.05	1.08
020125	2819	0.74	0.70	1.02
020128	2861	1.47	1.32	1.57
020129	2835	1.02	0.95	1.26
020130	7827	1.16	0.91	1.08
020131	2754	0.80	0.65	0.98

TAB. 8.13 – Taux de MHV obtenus sur les transcriptions manuelles des émissions de TF1 avec le vocabulaire de référence, le vocabulaire adapté selon la méthode ad hoc et le vocabulaire vectoriel

Jour	*#mots*	*Réf*	*Vocabulaire de référence*		*Vocabulaire adhoc*		*Vocabulaire vectoriel*	
			linéaire	sélective	linéaire	sélective	linéaire	sélective
02	2778	27.2	26.9	26.7	26.9	26.9	27.1	27.1
03	7828	28.0	28.0	27.9	27.6	27.4	27.4	27.1
04	2985	27.8	27.7	27.4	27.3	27.5	27.6	27.6
05	2849	26.4	25.5	25.5	25.0	25.0	25.2	25.0
07	2809	21.5	21.1	20.8	20.9	20.7	21.2	20.9
08	2983	23.9	23.9	23.8	23.6	23.6	23.6	23.6
09	2830	25.3	25.6	25.7	25.2	25.2	25.3	25.2
10	2771	26.8	26.7	26.7	26.4	26.3	26.7	26.4
11	2710	24.8	25.0	24.6	24.8	24.4	25.2	24.8
14	2867	20.4	20.3	20.4	20.1	20.0	20.2	20.3
15	8145	23.3	22.9	22.8	21.4	21.2	21.6	21.6
16	2813	21.9	22.0	21.9	22.1	22.2	22.3	22.5
17	2773	16.0	15.9	15.7	15.8	15.6	16.0	15.7
18	2753	20.3	20.1	19.7	19.8	19.9	20.1	20.2
21	2890	26.5	26.5	26.3	26.3	26.1	26.5	26.5
22	2853	22.1	21.8	21.9	21.7	21.8	21.5	21.7
23	2825	25.4	25.5	25.4	24.7	24.6	24.9	24.6
24	2846	27.9	27.3	27.4	27.4	27.4	27.4	27.7
25	2819	24.3	24.3	24.2	24.3	24.2	24.3	24.2
28	2861	27.8	28.0	28.1	27.8	27.9	28.2	28.3
29	2835	26.8	26.6	26.5	26.8	26.7	27.2	26.9
30	7827	21.4	21.5	21.4	21.4	21.2	21.4	21.4
31	2754	20.1	20.6	20.4	20.3	20.1	20.7	20.7

TAB. 8.14 – Résultats de reconnaissance obtenus sur le corpus de test télédiffusés, en faisant varier le vocabulaire (référence, adaptation ad hoc et vectoriel) et la technique d'adaptation des modèles de langage (interpolation linéaire et sélective). Les résultats du modèle de langage de référence sont également indiqué (Réf).

Table des figures

1 Schéma de principe des expérimentations. xix

2.1 Architecture type d'un système *SDR* 12

2.2 Exemple de courbe de rappel/précision. 14

3.1 Modèle source/canal du processus de production de la parole (figure extraite de [Gauvain and Lamel,2003] 22

3.2 Architecture d'un système de reconnaissance automatique de la parole 23

3.3 Modélisation d'un mot 25

3.4 Exemple de treillis 26

4.1 Loi de Zipf sur un corpus de journaux et un corpus de transcriptions manuelles . 34

4.2 Fréquences des fréquences des quadrigrammes observés sur les corpus de journaux et de transcriptions. 34

4.3 Fréquences des fréquences des quadrigrammes observés sur le corpus de journaux. 34

4.4 Processus de construction d'un modèle de langage 44

5.1 Taux de MHV en fonction de la perplexité du corpus de développement pour la segmentation *Années et type de source* obtenus lors des itérations de l'algorithme E.M. 71

5.2 Nombre de MHV en fonction des itérations de l'algorithme E.M.pour la segmentation *Années et type de source* (24 modèles sont interpolés) 72

6.1 Cadre expérimental pour l'évaluation des méthodes d'adaptation du vocabulaire. 76

6.2 % de MHV pour chaque jour j sur $T_{14}(j)$ avec le vocabulaire de référence et le vocabulaire adapté. 80

6.3 Nombre de mots nouveaux pour chaque jour 81

6.4 Comparaison du taux de MHV sur $T_{14}(j)$ obtenu avec 3 vocabulaires particuliers : $V_{adhoc}(j)$, $V_{adhoc}(01)$, $V_{adhoc}(31)$. 82

6.5 Comparaison des taux de MHV pour chaque jour j sur $T_{14}(j)$. 84

6.6 % de MHV pour chaque jour j sur $T_{14}(j)$ avec le vocabulaire de référence puis les 2 vocabulaires adaptés. 87

6.7 Comparaison du taux de MHV sur $T_{14}(j)$ obtenu avec 3 vocabulaires particuliers : $V_{vect}(j)$, $V_{vect}(01)$ et $V_{vect}(31)$. 88

6.8 Couverture lexicale sur $T_{14}(15)$ en fonction de la taille du vocabulaire pour les listes $\vec{Y_p}$ et $\vec{X_1}$ (std) 89

6.9 Évolution de la taille du vocabulaire nécessaire pour obtenir une couverture lexicale de 98,5% sur $T_{14}(j)$ avec les $\vec{Y_p}$ et $\vec{X_1}$ 90

8.1 Deux exemples de notices : une première notice concise, et une seconde plus riche. 107

8.2 Répartition des textes extraits du Web suivant les années. 109

8.3 Perplexité obtenue avec le modèle de référence et avec le meilleur modèle. 110

Liste des tableaux

1 Comparaison du contenu d'une notice documentaire et de la transcription automatique . . . xvii

2 Extraits de transcription d'émissions d'archive . . . xviii

2.1 Exemple de requêtes de test utilisées lors des campagnes d'évaluation *SDR* . . . 15

3.1 Exemple de transcription d'émission d'information américaine . . . 28

3.2 Exemple de transcription d'émission d'archive . . . 29

4.1 Architecture du modèle de langage connexioniste . . . 37

4.2 les dix quadrigrammes les plus fréquents observés sur un corpus de transcriptions d'émissions. . . . 39

4.3 Les dix quadrigrammes les plus fréquents observés sur un corpus de conversations téléphoniques . . . 39

4.4 Les dix quadrigrammes les plus fréquents observés sur un corpus de textes du Wall Street Journal. . . . 39

4.5 Caractérisation des sources de textes. . . . 45

4.6 Exemple de projection d'une phrase sur le vocabulaire. . . . 48

5.1 Caractérisation des sources de textes. . . . 61

5.2 Les mots hors du vocabulaire idéal observés sur le corpus de développement. . . . 62

5.3 Répartition des marqueurs morpho-syntaxiques des MHV du vocabulaire de référence observés sur le corpus de développement . . . 65

5.4 Évaluation des choix de segmentation de corpus en terme de taux de MHV mesurés sur le corpus de développement « Dev 2000 » et sur le corpus de test « Test 2002 » 69

5.5 Coefficients d'interpolation calculés via l'algorithme E.M et leurs corpus d'entraînement associés . . . 70

6.1 Moyenne, minimum et maximum du nombre de mots par jour et par source observés pour les mois de décembre 2001 et janvier 2002. 78

6.2 Les 20 premiers nouveaux mots de $V_{adhoc}(j)$ 82

6.3 Les 20 premiers nouveaux mots classés par leur rang lexical dans $V_{vect}(j)$ pour le mois de janvier 2002 87

7.1 Fréquences sur le corpus d'observation des catégories morpho-syntaxiques pour les mots hors du vocabulaire de référence. 95

7.2 Exemple de projection d'un texte d'entraînement sur le vocabulaire. 96

7.3 Proportion des étiquettes morpho-syntaxiques pour les 620 mots hors vocabulaire observés sur le corpus de test 97

7.4 Les résultats de reconnaissance, le taux de MHV obtenu avec le vocabulaire de référence, et le ratio entre le gain absolu sur le WER et le taux de MHV. 98

7.5 Exemple d'erreur de reconnaissance sur un extrait un mot nouveau « Izzo ». . . 99

7.6 Taux de reconnaissance sur les mots nouveaux. 100

7.7 Rang lexical obtenu par les formes de repli. 101

8.1 Nombre de documents et la durée par décennie pour le corpus EURODELPHES. . . 105

8.2 Nombre de mots, taux de mots hors vocabulaire, perplexité et taux d'erreur obtenus avec le système de référence modifié. 106

8.3 Perplexité par décennie estimée sur les données de test avec le modèle standard, le modèle d'époque et les modèles adaptés, puis λ_u. 109

8.4 Taux de mots hors vocabulaire obtenus sur le corpus de test RFI, avec le vocabulaire de référence, puis les deux vocabulaires adaptés, le premier selon la méthode *Ad hoc* et le second selon la méthode vectorielle (*Vect*). 114

8.5 Taux de mots hors vocabulaire obtenus avec les vocabulaires adaptés pour chaque jour via la méthode vectorielle, et calculés sur les transcriptions du corpus de test RFI. 114

8.6 Perplexité des transcriptions de référence obtenue avec différents modèles 115

8.7 Résultats de reconnaissance sur le corpus RFI 115

8.8 Répartition des mots nouveaux présents dans les transcriptions de référence, et pour chaque type la proportion correctement transcrite. 117

8.9 Résultats de reconnaissance obtenus sur le corpus de test télédiffusé 118

8.10 Résultats de reconnaissance obtenus sur le corpus EuroNews avec le vocabulaire et le modèle de référence, puis différentes configurations d'adaptation du modèle de langage. 120

8.11 Résultats de reconnaissance obtenus sur le sous-corpus de test EuroNews avec le système de référence et le système adapté (vocabulaire et modèle de langage). Les résultats sont exprimés en terme de taux de substitution et de WER. 120

8.12 Taux de mots hors vocabulaire obtenus avec les vocabulaires adaptés pour chaque jour via la méthode *ad hoc*, et calculés sur les transcriptions de référence 137

8.13 Taux de MHV obtenus sur les transcriptions manuelles des émissions de TF1 avec le vocabulaire de référence, le vocabulaire adapté selon la méthode ad hoc et le vocabulaire vectoriel 138

8.14 Résultats de reconnaissance obtenus sur le corpus de test télédiffusés 139