

Université Paris-Sud – Faculté des sciences d’Orsay
École Doctorale d’Informatique de Paris-Sud
Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur

Thèse

pour le diplôme de docteur en sciences, spécialité informatique,
présentée et défendue
le 11/04/2008 à Orsay (91)

par Thomas Pellegrini

Titre : Transcription automatique de langues peu dotées

Membres du jury :

Laurent Besacier (rapporteur)

Jim Glass (rapporteur)

Lori Lamel (directrice)

Isabel Trancoso (examinatrice)

Edouard Geoffrois (examineur)

Joseph Mariani (examineur)

Table des matières

Introduction	6
1 Présentation de la reconnaissance de la parole	10
1.1 Principe général des systèmes de reconnaissance	11
1.2 Modélisation de la syntaxe	12
1.3 Modélisation des prononciations	15
1.4 Modélisation acoustique	17
1.5 Décodage	20
1.6 Évaluation	21
1.7 Aperçu de l'état de l'art pour les langues bien dotées	22
1.8 Les langues peu dotées	23
1.9 Conclusion	33
2 Reconnaissance automatique de l'amharique	34
2.1 Présentation de la langue amharique	34
2.2 La langue amharique, une langue peu dotée?	36
2.3 Les données audio et textes	38
2.4 L'étape de normalisation	39
2.5 Le lexique de prononciation	40
2.6 Génération de prononciations à l'aide d'alignements phonotactiques	42
2.7 Modèles de phones amhariques	47
2.8 Modélisation linguistique	48

2.9	Évolution des performances en fonction de la quantité de données	49
2.10	Vaut-il mieux collecter des textes ou transcrire quelques heures de données audio ?	50
2.11	Premières expériences de décomposition des mots	61
2.12	Conclusion	63
3	Modélisation lexicale	65
3.1	Diversité des systèmes d'écriture	65
3.2	Translittération versus transcription	67
3.3	Conversion graphème-phonème	69
3.4	Quelques éléments de morphologie	70
3.5	Le mot : unité lexicale de base en reconnaissance automatique	72
3.6	Les sous-unités : une alternative ?	73
3.7	Segmentation morphologique	75
3.8	Différentes approches et résultats dans la littérature	77
3.9	Remarque sur la recombinaison des morphes	83
3.10	Conclusion	84
4	Sélection automatique de sous-unités lexicales	85
4.1	Présentation de l'algorithme <i>Morfessor</i>	86
4.1.1	Cadre mathématique de <i>Morfessor</i>	87
4.1.2	Décompositions d'un nouveau lexique à partir d'un modèle	90
4.2	Nouvelles propriétés et modifications apportées	90
4.2.1	Algorithme inspiré de Harris	91
4.2.2	Taille des morphes	93
4.2.3	Propriété fondée sur les traits distinctifs	93
4.2.4	Contraintes introduites dans la décomposition	95
4.3	Implémentation de <i>Morfessor</i>	96
4.3.1	Coût total	96
4.3.2	Probabilité de Harris	97
4.3.3	Probabilité de nombre d'occurrences	97

4.3.4	Exemple d' <i>illustration</i>	98
4.4	Conclusion	102
5	Sélection automatique appliquée à l'<i>amharique</i>	104
5.1	Les différents systèmes comparés	104
5.2	Les modèles de langage	106
5.3	Comparaison des performances des systèmes	109
5.3.1	Expérience complémentaire avec un ML d' <i>ordre plus élevé</i>	115
5.4	Conclusion	116
6	Sélection automatique appliquée au turc	119
6.1	La langue turque, une langue peu dotée?	119
6.2	Les données audio et textes	121
6.3	Expériences sur le corpus de parole lue	122
6.4	Expériences sur un corpus de parole broadcast news	130
6.5	Conclusion	134
	Conclusion et perspectives	136

Remerciements

Cette recherche a été conduite au LIMSI-CNRS au sein du groupe Traitement du Langage Parlé, qui offre des conditions scientifiques et matérielles tout à fait exceptionnelles. J’y ai été accueilli par Jean-Luc Gauvain, le directeur du groupe, que je remercie pour m’avoir fait confiance, et laissé totalement libre de mes choix d’investigation.

Je remercie très chaleureusement ma directrice de thèse, Lori Lamel, que je recommande vivement aux futurs doctorants potentiels en traitement de la parole. Elle a toujours veillé à bien me diriger dans les directions de recherche que je prenais, avec un enthousiasme pour la recherche très communicatif et un professionnalisme exceptionnel. Je remercie très fortement tout le groupe, j’ai eu des échanges avec pratiquement tout le monde, en particulier avec Martine Adda-Decker, qui m’a beaucoup inspiré grâce à ses travaux sur l’allemand entre autres, et qui a fortement contribué à la rédaction de ce mémoire, tant sur la forme que le fond.

Je remercie mes deux colocataires de bureau, Philippe Boula-de-Mareüil et Hélène Bonneau-Maynard, pour leur aide et leur amitié, Sophie Rosset qui m’a beaucoup aidé pour démarrer le travail sur l’amharique, Gilles Adda, Éric Bilinski, Olivier Galibert, Claude Barras, et tous ceux qui m’ont aidé un jour ou un autre.

Je n’oublie pas toute l’équipe de thésards du groupe, avec qui les repas au CESFO et les pauses cafés chez Nédé ont fortement contribué à renouveler l’énergie nécessaire pour persévérer. Je pense à Marc Ferras, Bianca Dimilescu, Cécile Woehrling, Laurence Vidrascu, Daniel Déchelotte et tous les autres.

Enfin je dédie ce mémoire à ma famille, à mes parents spécialement, ainsi qu’à mes frères et soeurs Chloé et Romain également, pour leur soutien et leur intérêt à me voir concrétiser ce travail. Pour conclure, je souhaite remercier tous mes amis et amies, qui font de ma vie à Paris un vrai bonheur, et spécialement Flavia, mais aussi les groupes de musique Untchak Attak et Roda do Cavaco.

Introduction

En 2007, le cabinet de recherche Computer Industry Almanac¹ estimait le nombre d'ordinateurs personnels en circulation dans le monde à un milliard de machines, soit un ordinateur pour 6,6 personnes en moyenne. Ce chiffre est accompagné de tendances qui montrent que la progression de l'informatisation des foyers diminue légèrement dans les pays dits développés, et augmente très fortement depuis quelques années dans les pays en voie de développement, nouveaux moteurs de la croissance de ce marché. Dès lors, le développement rapide de technologies de traitement numérique du langage, dans les langues de ces pays, est un enjeu essentiel.

La reconnaissance vocale, qui consiste en la conversion d'un signal audio de parole en une séquence de mots, est l'une de ces technologies, dont les applications sont très nombreuses. Il existe des applications directes, comme l'indexation automatique de documents audiovisuels, les systèmes de dialogue, ou simplement les logiciels de dictée. Les séquences de mots reconnus peuvent également servir d'entrée à d'autres systèmes, par exemple pour la traduction automatique de la parole, domaine en plein essor actuellement. Ces technologies, qui sont très avancées, sont réservées, pour l'instant, à un très petit nombre de langues. Il s'agit des langues des pays dits développés, ou de langues qui suscitent un intérêt économique ou politique, comme par exemple l'anglais, le français, l'arabe classique, le mandarin, le japonais, l'allemand, l'espagnol et le portugais, entre autres. Qu'en est-il de la très grande majorité des autres langues ?

La plupart des systèmes qui ont des applications réelles de nos jours, sont fondés sur une modélisation statistique de la parole. Il en résulte que de très grands corpus de données sont requis pour construire des modèles performants. Les progrès considérables qui ont été réalisés depuis les années 1990, ont permis l'émergence de recherches et de projets nombreux sur l'adaptation rapide des systèmes à des langues qui ne disposeraient pas, a priori, de quantités de données suffisantes. De nombreux termes plus ou moins équivalents existent dans la littérature pour désigner ces langues, qui sont pour certaines, parlées par des millions de personnes, mais qui ne disposent pas d'une activité et de ressources numériques importantes. Une qualification semble être plus utilisée néanmoins, il s'agit de l'expression « langues peu dotées » en français, et « less-represented languages » en anglais. Le projet actuel SPICE (Speech Processing : Interactive Creation and Evalua-

¹[http ://www.c-i-a.com](http://www.c-i-a.com)

tion Toolkit), par exemple, de l'université Carnegie Mellon, s'est intéressé entre autres à l'afrikaans, au bulgare, au vietnamien, à l'hindi, au konkani, au telugu et au turc [Schultz *et al.*, 2007]. Des projets plus anciens visaient à collecter des ressources pour des langues peu dotées, comme par exemple le projet Babel sur cinq langues est-européennes (bulgare, estonien, hongrois, roumain et polonais) [Roach *et al.*, 1996] et le projet GlobalPhone sur une quinzaine de langues avec des corpus de textes et de parole lue [Schultz, 2002]. D'une manière générale, pour les langues peu dotées, les technologies vocales ne sont peut-être pas la première lacune à combler, néanmoins la recherche et le développement sur ce thème, génère des outils et des corpus qui peuvent servir à d'autres tâches.

Les trois ressources nécessaires au développement d'un système de reconnaissance de la parole sont des corpus de textes comprenant au minimum quelques millions de mots, quelques dizaines d'heures de parole transcrites manuellement, et un lexique de prononciations qui fait le lien entre la modélisation acoustique et la modélisation linguistique de la langue étudiée. Les projets cités précédemment, ainsi que des travaux récents, comme par exemple les thèses « reconnaissance automatique de la parole pour des langues peu dotées » [Le, 2006] et « Sauvegarde du patrimoine oral africain : conception de système de transcription automatique de langues peu dotées pour l'indexation des archives audio » [Nimaan, 2007], ont principalement cherché à limiter le temps et les moyens nécessaires à la constitution des corpus d'apprentissage audio et textes, et ont mis l'accent sur la modélisation acoustique en étudiant la portabilité rapide des modèles acoustiques d'une langue vers une autre. Si cette thèse s'inscrit dans la lignée de ces travaux, nous avons préféré concentrer nos efforts sur les problèmes posés par le manque de textes, qui est, à nos yeux, le facteur le plus limitant pour les langues peu dotées. La constitution d'un corpus de textes peut être très difficile pour la très grande majorité de langues peu dotées, qui disposent d'une présence sur Internet très limitée. Pour tenter de pallier les problèmes qui en découlent, à savoir des taux de mots inconnus très élevés et des modèles de langage peu fiables, nous avons axé nos recherches principalement sur la modélisation lexicale, et en particulier sur la sélection des unités lexicales, mots et sous-unités, utilisés par les systèmes de reconnaissance.

Au cours de cette thèse, nous nous sommes attachés à toujours évaluer de manière quantitative les performances des méthodes proposées. Nous avons cherché à valider nos idées sur deux langues, l'amharique et le turc, sur des corpus de parole provenant d'émissions d'information de radio et de télévision (« broadcast news ») dans la mesure du possible, pour se confronter à des données réelles, du même type que celles qui sont utilisées dans les évaluations internationales des systèmes à l'état de l'art.

Ce mémoire est divisé en six chapitres. Après une brève présentation du principe général de la reconnaissance automatique de la parole par modèles statistiques, le premier chapitre définit les langues peu dotées en identifiant les problèmes posés par le manque de ressources numériques. En particulier, l'accent est mis sur le problème du manque de textes qui caractérise ces langues. Pour illustrer les difficultés rencontrées lors de l'élaboration d'un système de transcription pour une langue peu dotée, le chapitre deux

décrit les différentes étapes de développement d'un système pour une langue pour laquelle nous disposions de peu de ressources numériques, mais également de peu d'expertise linguistique, l'amharique, langue officielle de l'Éthiopie. Le corpus audio de parole contient 37 heures d'émissions d'information transcrites manuellement, et le corpus de textes collectés sur Internet totalise 4,6 millions de mots. L'un des premiers problèmes rencontrés fut la création du lexique de prononciations. L'approche la plus simple est l'approche graphémique, qui consiste à associer un phone à chaque graphème. Nous proposons une méthodologie originale de génération de variantes de prononciations à partir de l'approche graphémique, à l'aide d'alignements phonotactiques. Nous étudions ensuite l'évolution des performances d'un système de reconnaissance standard en fonction du nombre d'heures d'apprentissage des modèles acoustiques, en faisant également varier la quantité de textes utilisés pour estimer les modèles de langage. En fin de chapitre, les premières expériences de décomposition de mots sont décrites. Les légers gains obtenus par rapport à un système fondé sur les mots entiers ont été le point de départ du principal axe de recherche de cette thèse, à savoir la recherche automatique d'unités lexicales de reconnaissance pour optimiser la quantité de textes disponible.

Le chapitre trois présente différentes représentations lexicales pouvant être utilisées pour la reconnaissance de la parole, et dresse un état de l'art sur ce thème. Plusieurs méthodes sont décrites dans la littérature, différant de par l'usage qui est fait des sous-unités dans les modélisations acoustique et linguistique. Nous justifions la méthode que nous avons retenue, qui consiste à utiliser les sous-unités à tous les niveaux de modélisation. Nous justifions également le choix d'un algorithme de découverte de frontières de morphèmes. Il s'agit de l'algorithme appelé Morfessor, développé à l'université de technologies d'Helsinki et distribué sous license GNU/GPL. Le chapitre suivant décrit le cadre mathématique de Morfessor, qui cherche à maximiser de manière itérative la probabilité d'un lexique, en essayant de décomposer les mots en unités plus petites mais plus fréquentes, appelées *morphes*. Dans l'algorithme de départ, les propriétés utilisées pour estimer cette probabilité sont liées uniquement à des propriétés graphémiques des entrées lexicales. Nous décrivons l'introduction de nouvelles propriétés, qui tentent de prendre en compte des caractéristiques d'ordre oral dans le choix des décompositions. Une contrainte empêchant la création d'un morphe qui risque de se substituer à un autre déjà présent dans le lexique lors d'une transcription, une notion de distance acoustico-phonétique entre les phones qui composent les sous-unités, fondée sur des traits distinctifs, ont été introduites. D'autres modifications de l'algorithme ont été réalisées, en particulier l'introduction d'une probabilité inspirée de l'algorithme de Harris (1955), pour détecter les frontières de morphes de manière plus efficace que l'algorithme de départ.

Les deux derniers chapitres illustrent l'application de la sélection automatique d'unités de reconnaissance par des expériences de reconnaissance, sur deux langues : l'amharique et le turc. Les performances des systèmes de reconnaissance, mesurées en taux d'erreurs de mots, sont systématiquement comparées à celles des systèmes fondés sur les mots entiers. Les expériences sont menées sur des corpus de parole provenant d'émissions de radio

et télévision transcrites manuellement, mais également sur un corpus supplémentaire de parole lue pour le turc. Des gains légers mais significatifs sont obtenus avec les deux langues, ce qui est encourageant quant à la généralisation possible de la méthode, et son utilisation pour d'autres langues.

Chapitre 1

Présentation de la reconnaissance de la parole

Dans ce premier chapitre, nous présentons les fondements de la reconnaissance automatique de la parole et les différentes composantes des systèmes standards. La notion de langue peu dotée sera précisée en fin de chapitre.

La reconnaissance automatique de la parole a pour but de convertir un signal audio de parole en une séquence de mots correspondant au message linguistique sous-jacent. Les mots reconnus peuvent être le résultat final attendu comme dans les applications suivantes : dictée automatique, routage dans un centre d'appel, commandes simples (par mots isolés). Ils peuvent également être une entrée pour un traitement de type traitement du langage naturel, par exemple pour les applications suivantes : traduction automatique, indexation de documents audio, portails d'information, interaction avec des agents virtuels.

Le domaine de la reconnaissance de la parole a bénéficié d'avancées scientifiques considérables ces dix dernières années. L'évolution des technologies avec des ordinateurs de plus en plus performants et rapides permet l'accès aux traitements en temps réel. Si les principes de base de la reconnaissance n'ont pas particulièrement évolué depuis la formulation statistique introduite en 1976, les systèmes se sont considérablement complexifiés et il est difficile d'expliquer leur fonctionnement de manière exhaustive. L'objectif de ce chapitre n'est donc pas de décrire un système de reconnaissance en détail, mais de donner une vue d'ensemble des différentes composantes pour pouvoir bien situer les travaux réalisés au cours de la thèse.

1.1 Principe général des systèmes de reconnaissance

La formulation classique du problème de la reconnaissance de la parole continue a été introduite il y a un peu plus de trente ans, au milieu des années 70 [Jelinek, 1976; Bahl *et al.*, 1976]. L'objectif de la reconnaissance de la parole est de trouver la séquence de mots $\hat{M}$ la plus probable étant donné un signal audio de parole S . Du point de vue statistique, la parole est supposée être générée par un modèle de langage qui fournit des estimations des probabilités d'une séquence de mots M , un modèle de prononciations qui fournit des séquences de phones H associées aux mots, et un modèle acoustique qui code le message M dans le signal sonore S . La figure 1.1 illustre ce modèle de génération statistique de la parole. M , H et S sont vues comme des variables aléatoires, de telle manière que le problème de la reconnaissance de la parole consiste à estimer les densités de probabilités de ces variables.

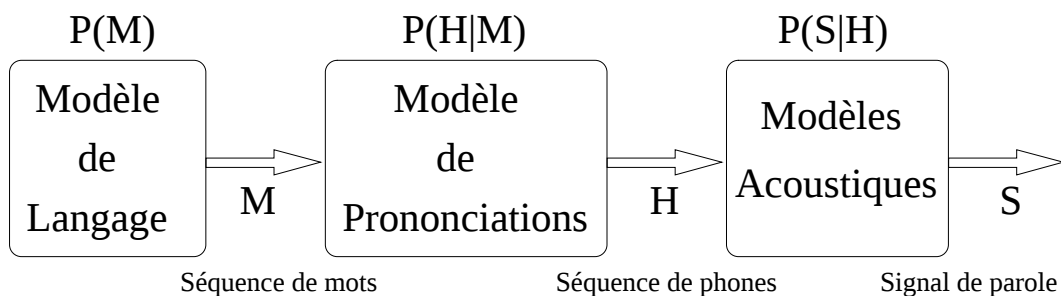


FIG. 1.1 – Modélisation statistique de la parole.

La recherche de $\hat{M}$ se traduit par la maximisation de $P(M|S)$, donnée dans l'équation 1.1.

$$\hat{M} = \operatorname{argmax}_M P(M|S) \quad (1.1)$$

Ne sachant pas évaluer directement la probabilité *a posteriori* $P(M|S)$, cette probabilité est décomposée de manière classique par la relation de Bayes, et devient l'équation 1.2.

$$\hat{M} = \operatorname{argmax}_M P(S|M)P(M)/P(S) \quad (1.2)$$

Où $P(S|M)$ est la probabilité conditionnelle, appelée « vraisemblance », d'observer le signal S à partir de la séquence de mots M . Le second terme $P(M)$ est la probabilité *a priori* d'avoir la séquence de mots M . Cette dernière probabilité est indépendante du signal audio, et met en jeu les propriétés syntaxiques de la langue. L'estimation de cette probabilité se fait à l'aide d'un modèle appelé modèle de langage. La probabilité du signal $P(S)$ ne dépend pas du message M , donc la maximisation sur les séquences de mots M ne fait pas intervenir $P(S)$. L'équation se simplifie en l'expression 1.3 :

$$\hat{M} = \operatorname{argmax}_M P(S|M)P(M) \quad (1.3)$$

La probabilité $P(S|M)$ peut être réécrite comme la somme des probabilités de prononcer le mot M avec une suite de phones H appelée « prononciation » : $\sum_H P(H|M)P(S|H)$. La somme est effectuée sur les différentes prononciations des mots de la séquence M .

Pour résumer, la formulation statistique de la reconnaissance de la parole aboutit à quatre considérations principales [Chou & Juang, 2003], qui sont développées dans les sections qui suivent :

1. Le problème de la modélisation du langage ou modélisation de la syntaxe, qui se traduit par l'estimation de la probabilité *a priori* $P(M)$.
2. Le problème de la modélisation des prononciations qui consiste à estimer $P(H|M)$. Cette modélisation est réalisée par un lexique qui donne pour chaque forme lexicale, la ou les prononciations associées. Ce lexique peut contenir des estimations des probabilités des prononciations, en particulier pour des langues qui présentent beaucoup de variantes comme l'arabe par exemple. L'utilisation de ces probabilités se généralise grâce à l'accès à des corpus de plus en plus grands.
3. Le problème de la modélisation acoustique, qui concerne le terme $P(S|H)$, qui représente le lien entre les suites de phones et les modèles acoustiques.
4. Le problème du décodage, qui consiste à rechercher la meilleure hypothèse $\hat{M}$. Les espaces de recherche sont particulièrement grands pour la reconnaissance de la parole à grands vocabulaires.

1.2 Modélisation de la syntaxe

Pour la reconnaissance de la parole continue, la seule information acoustique ne suffit pas pour transcrire les suites de mots correctement. La modélisation des suites de mots est indispensable, à la différence de la reconnaissance des mots isolés.

Dans l'équation 1.2, le terme $P(M)$ est la probabilité d'obtenir la séquence de mots M . La modélisation la plus couramment utilisée depuis de nombreuses années est celle des n -grammes. Elle consiste à donner la probabilité d'un mot à partir de la séquence de mots qui le précède. La probabilité d'une séquence de N mots est le produit des probabilités conditionnelles d'un mot sachant les mots qui précèdent (on parle d'historique) :

$$P(m_1 \dots m_N) = P(m_1) \prod_{i=2}^N P(m_i | m_1 \dots m_{i-1}) \quad (1.4)$$

Si l'on note m_1^{i-1} l'historique $m_1 \dots m_{i-1}$, l'équation s'écrit simplement :

$$P(m_1^N) = P(m_1) \prod_{i=2}^N P(m_i | m_1^{i-1}) \quad (1.5)$$

En pratique, la manipulation des probabilités de toutes les suites de mots possibles est irréalisable, et l'historique est tronqué à quelques mots. Les modèles trigrammes ou quadrigrammes, qui correspondent respectivement à un historique de deux mots et de trois mots sont les plus utilisés. Des mots balise, qui représentent les débuts et fins de phrases, sont de manière habituelle ajoutés automatiquement lors de la création des modèles de langage. Souvent notés '<s>' et '</s>', elles sont prises en compte dans l'historique des mots¹. Pour un ordre égal à trois, soit un modèle dit *trigramme*, l'équation 1.5 devient :

$$P(m_1^N) = P(m_1)P(m_2|m_1) \prod_{i=3}^N P(m_i|m_{i-2}^{i-1}) \quad (1.6)$$

Les probabilités des n -grammes sont calculées à partir des fréquences des mots d'un corpus de textes. Même pour les langues bien dotées, les quantités disponibles de textes pour estimer les probabilités des trigrammes ou des quadrigrammes ne sont pas suffisantes a fortiori pour les n -grammes d'ordre plus élevé. De nombreuses techniques de lissage ont été inventées pour pallier ce problème. Le lissage consiste à prendre de la masse de probabilité des n -grammes observés, pour donner une valeur non-nulle aux probabilités des n -grammes non-observés ou peu observés. L'une des techniques de lissage la plus utilisée est la technique dite de Kneser-Ney modifiée [Kneser & Ney, 1995]. Avec cette technique, les probabilités des n -grammes peu observés sont estimées comme avec les autres techniques de lissage, en faisant un repliement (« backoff ») sur un historique d'ordre moins grand. Pour un trigramme par exemple, le bigramme puis l'unigramme si nécessaire sont utilisés. L'originalité de la technique Kneser-Ney modifiée est de ne pas prendre la même distribution de probabilité pour les ordres plus petits que n . Au lieu de prendre la fréquence de l'historique d'ordre $n-1$ à savoir m_{i-n+1}^{i-1} , c'est le nombre de contextes différents dans lesquels se produit m_{i-n+1}^{i-1} qui est consulté. L'idée est que si ce nombre est faible alors la probabilité accordée au modèle d'ordre $(n-1)$ doit être petite et ce, même si m_{i-n+1}^{i-1} est fréquent. Ainsi le biais potentiel introduit par la fréquence de l'historique est évité.

En pratique, pour construire les modèles de langage (ML), nous avons utilisé la librairie SRILM [Stolcke, 2002]. Il existe cependant d'autres boîtes à outils, comme par exemple la « CMU SLM », pour « Carnegie Mellon Statistical Language Modeling (CMU SLM) Toolkit »².

¹Remarque : ces balises sont intégrées au lexique de reconnaissance, et sont associées au phone silence.

²Lien CMU SLM : http://www.speech.cs.cmu.edu/SLM_info.html

Évaluation des modèles de langage

Une question primordiale est de savoir comment deux modèles de langage peuvent être comparés en termes de performances d'un système de reconnaissance global. La façon correcte de procéder consiste à incorporer chaque modèle dans un système complet et d'évaluer quelle est la meilleure transcription en sortie du système.

Cependant, construire et tester un système complet prend beaucoup de temps, des heures voire des jours. Il existe cependant des mesures qui permettent de prévoir, avec plus ou moins de succès, l'efficacité d'un modèle de langage. Les plus connues et les plus utilisées sont la vraisemblance et la perplexité.

Les modèles de langage, introduits au paragraphe précédant, sont issus de concepts de la théorie de l'information. Un langage est vu comme une source discrète d'informations pouvant générer des séquences de mots comme par exemple $m_1^N = m_1 \dots m_N$. Cette conception a été introduite par C.E. Shannon dans l'article de 1948, qui donne une formulation mathématique de la théorie de la communication [Shannon, 1948].

Une manière de mesurer la qualité d'un modèle de langage est d'estimer la probabilité de séquences de mots qui ne font pas partie du corpus d'apprentissage du modèle. La probabilité d'un texte $M = m_1 m_2 \dots m_N$, appelée « vraisemblance » en français, « likelihood » en anglais et notée lh , est donnée par l'équation 1.7. Plus la vraisemblance est grande, plus le modèle est capable de prédire les mots contenus dans le corpus. Le chapeau de $\hat{P}$ est là pour rappeler que nous ne pouvons qu'estimer cette probabilité par des modèles (les modèles n -grammes en général).

$$lh(M) = \hat{P}(m_1 m_2 \dots m_N) \quad (1.7)$$

La grandeur la plus utilisée pour caractériser les performances d'un modèle de langage est la perplexité, définie dans l'équation 1.8 et souvent notée pp . Elle est équivalente à la vraisemblance mais fait intervenir une normalisation sur le nombre de mots du corpus de test. À cause de l'inversion dans l'équation 1.8, plus la probabilité de la séquence de mots est grande, i.e. plus la vraisemblance est grande, plus la perplexité est petite. Ainsi maximiser la vraisemblance est équivalent à minimiser la perplexité.

$$pp = \hat{P}(m_1^N)^{-1/N} = 1/\hat{P}(m_1^N)^{1/N} \quad (1.8)$$

Une interprétation courante de la perplexité consiste à voir cette grandeur comme le facteur de branchement moyen pondéré d'une langue. Le facteur de branchement moyen d'une langue est le taux moyen de mots qui peuvent suivre un mot donné de manière équiprobable. Une perplexité de 200 signifie qu'en moyenne, chaque mot du texte sur lequel elle a été mesurée, peut être suivi par 200 mots distincts de manière équiprobable. Prenons l'exemple de la reconnaissance de chiffres. Le vocabulaire est un vocabulaire fermé de dix mots équiprobables de probabilité $1/10$. Pour une chaîne de N chiffres, la

perplexité vaut :

$$\text{pp} = \hat{P}(m_1^N)^{-1/N} = (1/10^N)^{-1/N} = (1/10)^{-1} = 10 \quad (1.9)$$

Dans le cas où les mots sont équiprobables, la perplexité est égale au facteur de branchement. En revanche, si l'un des chiffres est beaucoup plus fréquent que les autres alors on s'attend à une perplexité (ou facteur de branchement pondéré) plus petite, le facteur de branchement étant toujours égal à 10.

Enfin, voici deux remarques à prendre en considération lorsque l'on compare des modèles de langage :

- Une réduction de perplexité (ou une augmentation de vraisemblance) n'implique pas toujours un gain de performances d'un système de reconnaissance,
- En général, la perplexité de deux modèles n'est comparable que s'ils utilisent le même vocabulaire. Sinon, il faut utiliser une perplexité normalisée qui simule un nombre de mots identique.

Bien que des modèles de langage avec des mesures de perplexité qui diminuent tendent à améliorer les performances d'un système de reconnaissance, il existe dans la littérature des études qui reportent des diminutions importantes de perplexité n'ayant peu ou pas apporté de gain de performance [Martin *et al.*, 1997; Iyer *et al.*, 1997].

Étoffons le deuxième point qui est très important. La perplexité est une moyenne sur le nombre de mots du texte sur lequel elle est mesurée. Si l'on compare, par exemple, un modèle basé sur des mots entiers et un modèle basé sur des sous-unités, il faudra conserver la puissance $-1/N$ qui fait intervenir le nombre de mots N dans le calcul de la perplexité pour le modèle basé sur les sous-unités. Pour ce modèle, la perplexité prend la forme donnée dans l'équation 1.10, où M est le nombre de sous-unités qui composent le texte de test qui comporte N mots au départ.

$$\text{pp} = \hat{P}(m_1^M)^{-1/N} \quad (1.10)$$

1.3 Modélisation des prononciations

Pour décoder ou segmenter en phones un signal de parole, il est nécessaire de pouvoir associer aux formes orthographiques des mots, une forme phonémique qui donne la séquence de modèles acoustiques associée. En réalité, cette forme phonémique que l'on appelle « prononciation », n'est pas forcément unique, et il peut être nécessaire d'associer plusieurs prononciations possibles à une forme orthographique. On parlera de « prononciation principale » pour la prononciation « standard »³, et de « variantes de

³Il reste à définir quelle est la prononciation standard. Nous aborderons ce problème au cours des chapitres suivants, mais d'une manière générale, il s'agira de la prononciation la plus fréquente.

prononciation » pour les autres prononciations.

Le jeu de phones ou de phonèmes, qui représente les modèles acoustiques élémentaires (on dira « indépendants du contexte »), dépend de la langue étudiée. Des exemples de jeux de phones classiques comportent 45 unités pour l’anglais, 50 pour l’allemand et l’italien, 35 pour le français, 25 pour l’espagnol [Chou & Juang, 2003]. Plus de détails sur les modèles acoustiques suivront dans la section 1.4.

Le tableau 1.1 donne un extrait du dictionnaire de 65k mots utilisé au LIMSI pour le français.

évoquera	evokəɾa	evokɾa		
(...)				
événement	evɛnəmã	evɛnmã	evɛnəmât(V)	evɛnmât(V)

TAB. 1.1 – Deux exemples d’entrées du dictionnaire de 65k mots du LIMSI pour la transcription automatique du français.

Le mot « évoquera » a une variante de prononciation due au schwa [ə], qui est fréquemment éliée en français. De même, le mot « événement » possède plusieurs variantes de prononciation, celles qui finissent par un ‘t’ correspondent au cas où la liaison avec le mot suivant qui commence par une voyelle (contexte noté ‘(V)’) est réalisée. Cet exemple montre qu’a priori la connaissance de la langue étudiée est nécessaire pour établir la correspondance écrit/oral. Cependant, en général, les lexiques de prononciation utilisent une unique prononciation par entrée lexicale, et les variations de prononciation sont modélisées par les modèles acoustiques de phones en fonction de ce qui est effectivement observé dans le corpus d’apprentissage. La génération des prononciations, même sans variante, dépend des relations graphèmes/phonèmes de la langue en question. Pour les langues pour lesquelles la conversion graphème/phonème est relativement directe, la création d’un tel dictionnaire est simple. C’est le cas du français, de l’espagnol, du portugais, de l’italien, de l’arabe, du turc et de l’amharique par exemple. En revanche une langue comme l’anglais a beaucoup de graphèmes qui se prononcent différemment en fonction du contexte avec de nombreuses exceptions aux règles que l’on pourrait établir. Pour cette raison la création d’un lexique de prononciations pour cette langue, est principalement manuelle. Le problème de la création d’un lexique de prononciations sera repris plus en détail dans le chapitre 3.

La figure 1.2 est un exemple d’alignement d’un segment de phrase en amharique. Sont montrés le spectre du signal audio de parole, et les alignements de ce segment en phones (première ligne d’étiquettes) et en mots (deuxième ligne d’étiquettes). Le mot « mEto » est un exemple d’alignement où le système a utilisé une variante de prononciation, c’est-à-dire que ce n’est pas la prononciation principale, identique à la forme graphémique, qui a été utilisée. Il a été aligné avec le phone [u] à la place du phone [o]. On peut remarquer que certains schwas (représentés par [x]) sont éliés, les deux schwas du mot

« ?ErxEJEnxtina » et le premier schwa du mot « kE?Enxdx ». Les phones représentés par un point sont des silences insérés automatiquement par le système d’alignement.

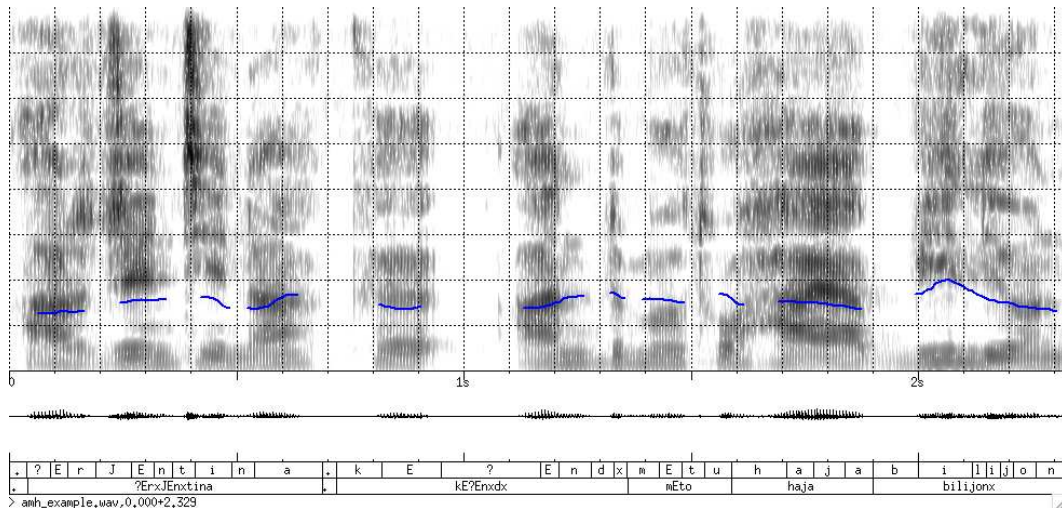


FIG. 1.2 – Extrait de spectrogramme avec segmentation en phones et en mots en amharique. Le mot « mEto » est aligné avec la variante [mEtu]. Certains schwas (notés ‘x’) sont élidés, comme ceux de « ?ErxEJEnxtina » par exemple.

Des alignements de corpus audio transcrits peuvent servir à identifier et sélectionner des prononciations. Pour la langue amharique par exemple, des variantes de prononciation ont été établies à partir d’alignements syllabo-tactiques [Pellegrini & Lamel, 2006]. Ces expériences seront décrites dans la section 2.6. Dans [Nakajima *et al.*, 2000], un corpus de transcriptions phonétiques est utilisé pour générer des variantes de prononciations phonologiques, utilisées sur des mots erronés lors de la reconnaissance automatique. L’utilisation de ces variantes montre de légers gains sur de la parole conversationnelle japonaise (8% relatif sur un taux d’erreurs de mots de référence d’environ 18%). La qualité d’un lexique de reconnaissance est primordiale pour obtenir un système global performant.

1.4 Modélisation acoustique

Modèles de phones, monophones, triphones

La plupart des systèmes actuels modélisent la parole continue par composition d’unités élémentaires appelées « phones ». Généralement, aucune hypothèse n’est faite sur le lien qui existe entre ces unités qui sont apprises sur le corpus de données, et les unités linguistiques que sont les phonèmes. Ces unités ne sont pas des phonèmes au sens linguistique du terme. Dans [Lee *et al.*, 1990] par exemple, les phones sont appelés « PLU » pour

« Phone-Like Units ». Ces phones de base sont déclinés en phones dits « hors contexte » ou « monophones », et en phones « en contexte ». Les phones en contexte sont appelés PIC en anglais, pour « phones in context ». Les monophones ne modélisent pas explicitement le contexte dans lequel ils apparaissent à la différence des phones en contexte. Les modèles en contexte sont en général des triphones, c'est-à-dire des phones qui dépendent des phones premiers voisins à gauche et à droite. Leur nombre est fonction principalement de la quantité de données audio d'apprentissage disponible, mais également de la langue étudiée, qui offre plus ou moins de contextes distincts.

Modèles de Markov Cachés

Les phones sont classiquement modélisés par des modèles de Markov cachés (MMC) de type gauche-droite qui sont bien adaptés pour représenter le flux temporel et la variabilité de la parole [Rabiner & Juang, 1986; Young, 1996]. Les MMC modélisent des vecteurs de paramètres acoustiques qui sont évalués par un pré-traitement du signal audio ou « acoustic front-end » en anglais. Les paramètres couramment utilisés sont des coefficients cepstraux obtenus par une analyse du logarithme du spectre espacé selon une échelle « Mel » ou « Bark » non-linéaire, qui est proche du traitement non-linéaire de l'oreille humaine [Davis & Mermelstein, 1980]. Le second type de paramètres très utilisés sont les paramètres PLP pour « Perceptual Linear Prediction », pour lesquels la transformation appliquée au logarithme du spectre est différente [Hermansky, 1990]. Les coefficients PLP sont légèrement plus robustes en présence de bruit de fond [Kershaw *et al.*, 1996].

Un modèle de phone peut tenir compte des phones voisins, auquel cas il sera appelé modèle dépendant du contexte. Il sera alors précisé s'il s'agit d'un diphone, d'un triphone, etc, en fonction du nombre de voisins pris en compte. Un modèle plus élémentaire ne modélise pas le contexte, et dans ce cas on parlera de monophone ou de modèle indépendant du contexte. La figure 1.3 montre un exemple de MMC triphone à trois états pour le phone $[a]$ des mots anglais « wand » ou « wan » également, dont les transcriptions phonétiques sont respectivement $[wand]$ et $[wan]$. Les MMC sont décrits par des probabilités de transition d'un état à un autre, notées a sur la figure, et par des densités de probabilités des observations, notées b , qui permettent d'associer une portion de signal au meilleur modèle de phone correspondant. De manière classique, les probabilités d'observation sont modélisées par des densités de probabilité de type mélange de gaussiennes. Il faut remarquer que des modèles à cinq états existent également, c'est le cas du système de BBN par exemple [Nguyen *et al.*, 2005].

Modèles à états liés

Les modèles acoustiques sont dits à états liés lorsque leurs états sont mis en commun. Cette opération est en général effectuée par un arbre de décision, et les états sont regroupés selon des critères linguistiques définis au préalable, il s'agit de traits phonétiques généraux. Les états liés partagent les mêmes jeux de gaussiennes mais également les poids dans les mélanges de gaussiennes. L'opération qui consiste à partager les états entre eux (« state-tying »), permet de regrouper des contextes peu observés [Bahl *et al.*, 1991;

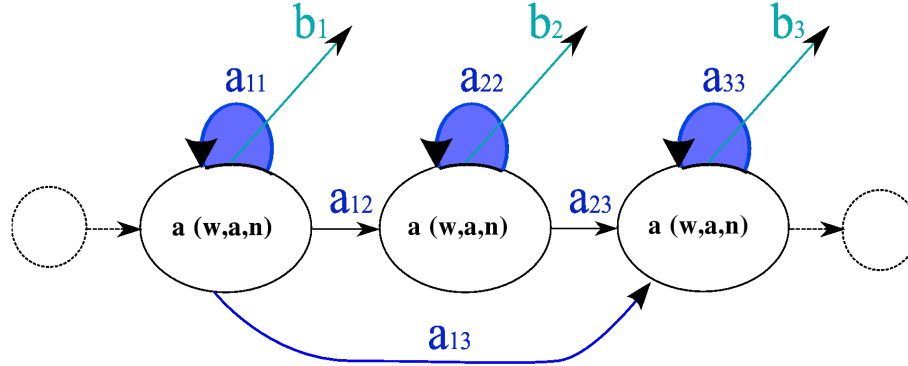


FIG. 1.3 – Exemple de modèle triphone à trois états pour le phone [a] des mots anglais « wand » ou « wan » transcrits en [wand] et [wan] respectivement. Le triphone est représenté par $a(w,a,n)$. Les probabilités de transition correspondent aux symboles a_{ij} , et les densités de probabilité d'observation aux symboles b_i . Sur cette figure, un saut entre l'état 1 et l'état 3 est possible.

Young *et al.*, 1994]. Dans le système pour l'anglais américain du LIMSI [Gauvain *et al.*, 2002], les contextes peu observés sont regroupés en regardant en premier lieu s'ils ont un contexte droit en commun (co-articulation régressive), sinon un contexte gauche en commun (co-articulation progressive), et enfin ceux pour lesquels aucun contexte en commun n'a été trouvé sont regroupés en modèles indépendants du contexte (on parle de backoff vers les diphtonges puis vers les monophones). Dans ce système, 30k contextes sont modélisés par 10k états liés. Le nombre de gaussiennes est de 16, 32 ou 64 pour le système broadcast news, et de 32 ou 64 pour les données de type CTS (Conversational Telephone Speech). Le nombre de contextes modélisés dépend principalement de la quantité de données d'apprentissage disponible, mais également de la langue traitée (la phonologie de la langue traitée offre plus ou moins de diversité de contextes), et du nombre de phones de base (parfois pour certaines expériences un jeu réduit de phones est utilisé, dans ce cas le nombre de contextes est plus petit qu'avec un jeu de phones plus grand).

Apprentissage des modèles

Les paramètres des MMC qui comprennent les probabilités de transition entre états, les moyennes, les variances et les poids des mélanges de gaussiennes, sont estimés sur des alignements de transcriptions de données audio d'apprentissage.

Au cours de l'opération appelée segmentation ou alignement, le signal audio est découpé en tronçons, associés chacun à un seul phone.

Pour réaliser les premiers alignements, plusieurs techniques sont utilisées, soit des tech-

niques de « flat start », soit des techniques de « bootstrap » qui utilisent des modèles pré-existants d'une ou plusieurs autres langues.

Approches « flat start »

L'approche communément appelée « flat start », est la technique la plus simple pour initialiser les paramètres des MMC. Elle consiste à mettre à zéro les probabilités de transition que l'on veut interdire, par exemple les transitions d'un état vers un état antérieur (modèles gauche-droite). Toutes les autres probabilités de transition entre états sont mises équiprobables. Pour les probabilités d'observation, les moyennes et les variances des gaussiennes sont toutes initialisées aux mêmes valeurs, à savoir la moyenne et la variance estimées sur toutes les données d'apprentissage.

Approches « bootstrap »

Deux approches principales dites de « bootstrap » existent pour initialiser les modèles acoustiques [Schultz & Waibel, 2001]. La première approche consiste à choisir des modèles acoustiques de systèmes de reconnaissance existants pour segmenter les données transcrites manuellement dans la langue cible. Cette méthode est fréquemment utilisée mais est rarement explicitement mentionnée dans la littérature. La deuxième approche consiste à prendre un jeu de modèles acoustiques multilingues génériques qui couvrent un grand nombre de phonèmes [Schultz & Waibel, 2001]. Cette dernière technique peut être utile pour réaliser un système pour une langue avec très peu de données, typiquement moins de 10h de transcriptions audio.

Chaque segment de parole est aligné soit de manière itérative avec l'algorithme de Baum-Welch [Baum *et al.*, 1970], qui prend en compte tous les chemins qui passent par un état, soit uniquement avec la meilleure séquence d'états possible (alignement de type Viterbi, ou alignement forcé) [Viterbi, 1967]. Après l'alignement, les paramètres des MMC sont estimés à l'aide d'une procédure EM (Expectation/Maximization) en partant d'une seule gaussienne par état, qui est divisée jusqu'à obtenir le nombre maximal de gaussiennes voulu, pris typiquement entre 8 et 64 gaussiennes [Gauvain *et al.*, 2002].

1.5 Décodage

Le principe général de la reconnaissance de la parole a été présenté dans la section 1.1. Le module qui effectue en pratique la reconnaissance est appelé décodeur. Son rôle est de chercher dans un espace d'hypothèses très grand, le meilleur chemin qui donnera la séquence de mots la plus probable.

En général, la première étape d'un décodeur consiste à identifier les parties du signal audio qui sont effectivement de la parole. Le problème du partitionnement des données audio ne sera pas abordé ici, une vision d'ensemble pourra être trouvée dans [Chou & Juang, 2003].

L'étape suivante est le décodage lui-même. Il existe de nombreuses stratégies de décodage,

et l'emploi de l'une ou l'autre dépend des contraintes que l'on se fixe : contrainte de temps réel et taille de vocabulaire entre autres. La stratégie dépend également fortement des contraintes matérielles, en termes de ressources informatiques (temps CPU et mémoire). Un point commun entre les différents systèmes est le compromis nécessaire entre la taille des modèles (en nombre de paramètres), et les réductions de l'espace de recherche (élagage ou « pruning » en anglais) [Chou & Juang, 2003]. Plus les modèles sont grands, plus les ressources informatiques doivent être importantes, mais dans ce cas, l'élagage peut être important et réduire le coût machine.

Les systèmes qui ont été construits au cours de cette thèse, l'ont été dans une perspective de recherche, et donc la contrainte temps réel n'a pas été prise en compte. La stratégie utilisée a été conservée pour tous les systèmes testés : il s'agit d'un décodeur classique à deux passes. La première passe génère des hypothèses qui servent à adapter de manière non-supervisée les modèles acoustiques aux données. Une deuxième passe génère un treillis de mots, par phrase, qui est une représentation compacte des hypothèses de mots, avec scores acoustiques et linguistiques. Les treillis sont ensuite « rescorés » par un modèle de langage, en général d'ordre plus élevé que les modèles utilisés lors des deux passes précédentes. Ce « rescoring » donne la phrase hypothèse la plus probable.

1.6 Évaluation

Les systèmes de reconnaissance sont traditionnellement évalués avec une mesure d'erreur moyenne sur les mots appelée WER, pour Word Error Rate. Il est défini par la somme de trois types d'erreurs qui sont l'insertion $\#I$, la substitution $\#S$ et l'élision $\#D$ de mots, moyennée par le nombre de mots N de la référence : $(\#I + \#S + \#D)/N$. Ce taux est calculé après alignement dynamique de l'hypothèse du décodeur avec une transcription de référence, à l'aide d'un calcul de distance d'édition minimale entre mots. Le résultat sera le nombre minimal d'insertions, de substitutions et d'élisions de mots, pour pouvoir faire correspondre les séquences de mots de l'hypothèse et de la référence. D'après sa définition, le WER peut être supérieur à 100% à cause des insertions.

La boîte à outils SCTK⁴ (« Scoring Toolkit ») du National Institute of Standards and Technologies (NIST) fournit le programme *scite* pour aligner les hypothèses et les références, calculer les WER et faire des analyses fines des erreurs. *scite* peut fournir des informations très utiles comme les mots les plus substitués, insérés ou élidés, des taux d'erreurs par locuteur peuvent être obtenus (si les segments possèdent une étiquette de locuteur). Voici un exemple d'alignement de la sortie d'un décodeur (HYP) d'une émission d'information de la radio France Inter, avec la référence (REF) :

L'hypothèse contient une substitution et une insertion :

$$\text{Word Error Rate} = 100(1+1)/9=22,2\%$$

⁴téléchargeable à l'adresse <http://www.nist.gov/speech/tools/index.htm>

REF :	pas question	*****	RÉPONDENT	l' union européenne et la russie
HYP :	pas question	RÉPOND	DE	l' union européenne et la russie
Eval :		I	S	

TAB. 1.2 – *Exemple d'alignement d'une phrase, sortie de scilite. Émission de radio (France inter 2003).*

Enfin, SCTK fournit également un programme `sc_stats` qui réalise des tests de significativité statistique, notamment le test « Matched-Pair Sentence Segment Word Error » (MAPSSWE) [Pallett, 1990]. Ce programme sera utilisé dans la section 5.3 pour vérifier que les différences de performances des systèmes sont significatives.

1.7 Aperçu de l'état de l'art pour les langues bien dotées

Le tableau 1.3 résume les taux d'erreurs moyens de systèmes à l'état de l'art pour différentes tâches, pour l'anglais américain. Il s'agit de taux d'erreurs indicatifs qui peuvent varier beaucoup selon les conditions acoustiques, et qui ont été obtenus dans certaines conditions d'apprentissage et de test. Par exemple, ajouter un bruit de voiture avec un rapport signal sur bruit de 10dB, peut multiplier par deux jusqu'à quatre le taux d'erreur. Un accent prononcé de non-natif peut également diminuer les performances considérablement. Une étude sur des locuteurs japonais anglophones lisant des textes en anglais (de type information ou Broadcast news) reporte des taux d'erreurs autour de 65% et plus [Tomokiyo, 1991].

Les chiffres qui sont donnés ici concernent la reconnaissance de la parole continue à grand vocabulaire (taille de quelques dizaines de milliers de mots). Les types de parole sont donnés par ordre décroissant de performances.

Parole lue

La parole lue est considérée plus simple à reconnaître parce qu'elle est contrôlée, bien articulée et enregistrée dans des conditions acoustiques favorables. La dernière compétition DARPA sur de la parole lue, qui a donné des WER autour de 7%, a eu lieu en 1995/1996, et depuis quelques années la parole lue n'est en général plus le type de parole sur lequel travaille la communauté, en ce qui concerne les langues bien dotées. Ce taux d'erreurs avec des systèmes actuels serait sans doute peu différent à quantité de données d'apprentissage égales.

Parole « broadcast news »

La parole « broadcast news », issue d'émissions radio-télévision a été et reste l'objet de recherches actives en raison des applications très intéressantes d'archivage, d'indexation automatique de grandes quantités de documents audio et vidéo. À la différence de la parole

lue, la parole « broadcast news » est une parole plus ou moins contrôlée, qui provient de situations réelles qui ne sont pas au départ liées à la transcription automatique. Le taux d’erreurs de mots indicatif de 10% est celui d’un système qui date de 2004, qui est la combinaison de systèmes de deux laboratoires BBN et LIMSI-CNRS. Les modèles acoustiques ont été entraînés sur plusieurs milliers d’heures de données audio de parole transcrites manuellement et automatiquement, et les modèles de langage ont été estimés sur des corpus de textes totalisant plus d’un milliard de mots [Nguyen *et al.*, 2004]. Plus précisément, deux types de parole broadcast news sont distingués, la parole broadcast news et la parole broadcast news conversationnelle, issue par exemple d’entrevues au cours d’une émission. Sur ce dernier type de parole, les taux d’erreurs sont d’environ 25% sur l’anglais américain.

Parole conversationnelle

Enfin, le type de parole considéré comme le plus difficile à transcrire est la parole conversationnelle, qui représente un défi actuel. La parole conversationnelle est la parole la moins contrôlée, la plus difficile à modéliser. Les taux d’erreurs de mots variant entre 30% et 40% en transcription de conversations téléphoniques sur les corpus Switchboard et multilingual Callhome (en espagnol, arabe, mandarin, japonais, allemand), témoignent de cette difficulté [Young & Chase, 1998]. Pour l’anglais américain, les chiffres actuels sont meilleurs, proches de 25% d’erreurs de mots.

<i>Tâche</i>	<i>Vocabulaire(# mots)</i>	<i>WER (%)</i>
Parole lue ^a	65k	7
Broadcast news ^b	65k	10
Parole conversationnelle ^c	65k	35
Parole conversationnelle (anglais américain)	65k	25

TAB. 1.3 – *Taux d’erreurs de mots moyens de systèmes de reconnaissance à l’état de l’art pour les trois types de parole habituellement utilisés : parole lue, Broadcast news et conversationnel.*

^aNorth American Business News Task 1995/1996 (DARPA benchmark test)

^bDonnées de type Broadcast news sans restriction

^cParole conversationnelle au téléphone

1.8 Les langues peu dotées

En laissant de côté les problèmes liés aux critères qui définissent une langue et en particulier la délicate distinction entre langue et dialecte, le nombre de langues dans le monde est en général estimé entre 5000 et 6000 langues [Amorrrortu *et al.*, 2004]. La distribution géographique des langues est très inégale selon les continents. Pour un total estimé à 6000 langues, presque deux tiers proviennent des continents africains et asiatiques (un tiers pour chaque continent), alors que seulement 3% sont des langues européennes. Enfin,

les langues des continents américains et de la zone pacifique représentent respectivement 15% et 18% des langues du monde [Grimes, 1996-2000]. Selon [Crystal, 2000], 82% des langues du monde ont moins de 100k locuteurs, et 56% moins de 10k locuteurs. Un faible nombre de locuteurs n'est pas le facteur unique déterminant le rayonnement d'une langue, néanmoins ces pourcentages montrent qu'une majorité de langues risquent de disparaître au profit d'autres langues dominantes [Hagège, 2000]. Pour des institutions comme l'UNESCO, le développement de ressources et d'outils numériques pour de telles langues est une étape nécessaire pour tenter de préserver la diversité linguistique menacée [Hagège, 2005].

Dans ce chapitre, nous allons définir ce que sont les langues peu dotées dans le contexte du traitement de la parole, et plus précisément de la reconnaissance de la parole. Le manque de données d'apprentissage pose le problème de l'estimation des modèles probabilistes utilisés pour la reconnaissance de la parole. Nous allons tenter de montrer que le manque de textes est le facteur le plus limitant lors de l'élaboration d'un système pour une langue peu dotée. Puis nous discuterons les études qui existent dans la littérature sur le lien entre les quantités de données d'apprentissage et les performances des systèmes.

Définition d'une langue peu dotée

Seules quelques dizaines de langues disposent d'une diffusion très large sur Internet. La langue la plus utilisée est bien sûr l'anglais. En 2006, le moteur de recherche Google indexait un total de 8 milliards de pages, dont une très large majorité de pages sont anglophones. La figure 1.4 donne la proportion de langues sur le Web à partir d'un échantillon de pages prises au hasard. Cette figure est issue d'une étude réalisée par l'OCLC (Online Computer Library Center) [Lavoie & O'Neill, 1998-2003] : des adresses de protocole Internet (IP) ont été générées aléatoirement, et lorsqu'elles correspondaient effectivement à un site Web, les auteurs téléchargeaient la page d'accueil et activaient un système d'identification automatique de langue [O'Neill *et al.*, 1997]. 72% des pages identifiées sont anglophones. Mis à part le japonais (3%), le chinois (2%) et le russe (1%), toutes les langues identifiées sont des langues occidentales (anglais, allemand, français, espagnol, etc). La catégorie « Autres » qui correspond à des langues qui n'ont pas été identifiées, représente seulement 2%. Cette étude montre très bien l'importance de l'anglais, suivi de très loin par une petite dizaine de langues. Cependant, elle a été réalisée en 2003 et le Web évolue de plus en plus vite. Dans le rapport d'un projet pluridisciplinaire de l'UNESCO « Initiative B@bel » [Wright, 1998-2003], destiné à trouver les moyens de protéger les langues de faible diffusion dans le monde, l'équipe de recherche affirme que « la tendance à la diversité linguistique va nécessairement s'accroître, à mesure que l'accès et la participation aux échanges sur Internet se développeront dans toutes les régions du monde ». Des études plus récentes montrent que l'anglais est de plus en plus concurrencé, par des langues latines (espagnol, français, portugais), mais surtout par le chinois et l'arabe, comme le montre la figure 1.5, qui donne le nombre d'utilisateurs

d'Internet en millions pour une dizaine de langues, selon une enquête qui date du 30 novembre 2007, disponible sur le site « Internet World Stats » (internetworldstats.com).

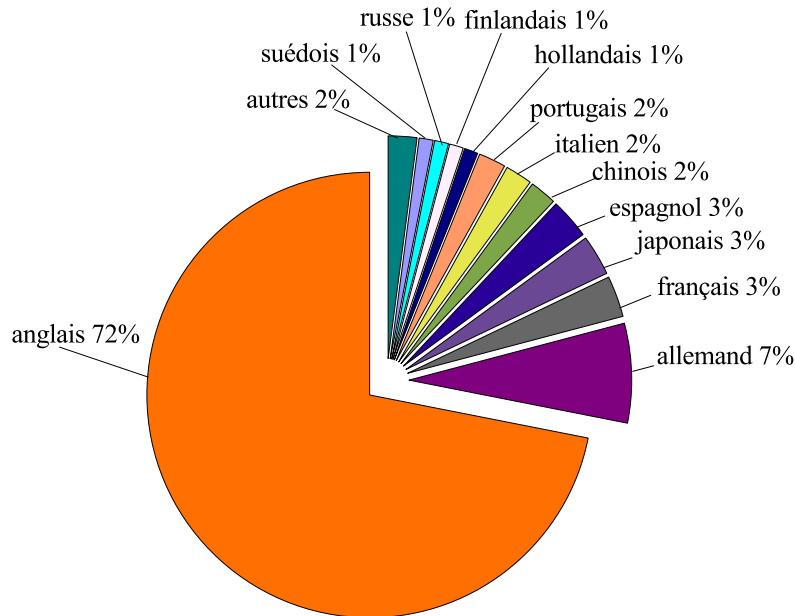


FIG. 1.4 – Proportions de différentes langues sur le Web à partir d'un échantillon au hasard de pages Web [Lavoie & O'Neill, 1998-2003]. Cette étude date de 2003.

Pour compléter ces chiffres, une autre étude est disponible sur le même site (internet-worldstats.com), qui donne l'évolution en pourcentage du nombre d'utilisateurs d'Internet par langue, entre 2000 et 2007. La figure 1.6 représente cette évolution, montrant entre autres que c'est l'arabe qui a eu la plus forte augmentation, avec plus de 1500% de progression. Viennent ensuite le portugais, la catégorie « autres », qui regroupe des langues non-identifiées, moins classiques, puis le chinois avec des progressions autour de 500%. La troisième position de la catégorie « autres » justifie en quelque sorte l'intérêt que l'on peut porter à d'autres langues que celles habituellement étudiées dans les technologies de la langue, et en particulier le traitement de la parole, et justifie le sujet de cette thèse, qui est un sujet tout à fait d'actualité.

Le fait qu'une langue soit déclarée langue officielle, ou qu'elle soit parlée par un grand nombre de locuteurs, n'implique pas forcément une présence importante de cette langue dans les médias et en particulier sur Internet. Une langue officielle comme le Quechua, parlée par 10 millions de personnes, n'est pratiquement pas présente sur le Web. À

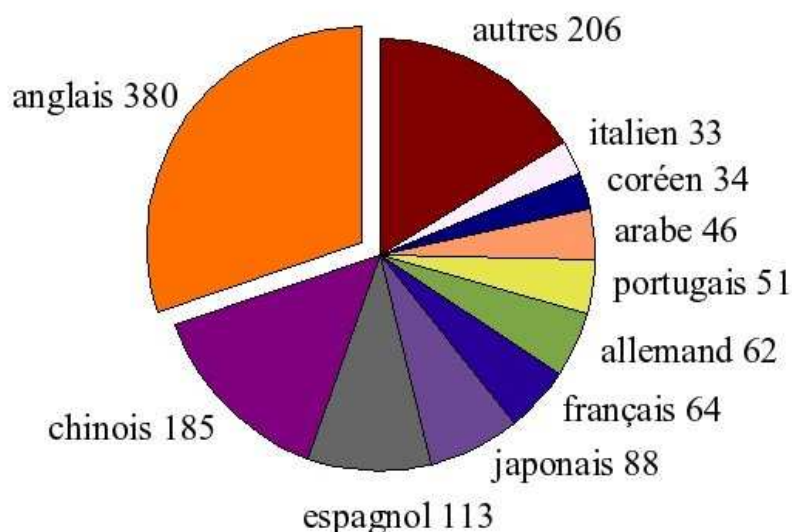


FIG. 1.5 – Nombre d'utilisateurs d'Internet par langue, en millions. Cette étude date du 30 novembre 2007. Source : *internetworldstats.com*

l'inverse, l'islandais, parlé par 300k personnes, est très dynamique et très utilisé dans les médias et sur Internet.

La figure 1.7 montre le nombre d'utilisateurs d'Internet par pays, pris sur des échantillons de 1000 personnes. Les pays sont classés en abscisse par indice de développement humain (IDH) décroissant, ou rang IDH croissant. Cette figure est tirée de chiffres donnés dans le rapport annuel de l'UNESCO sur le développement humain [UNDP, 2006].

Plus l'indice IDH est faible, plus le rang IDH est grand. Cet indice est l'indicateur habituel pour évaluer le niveau de développement d'un pays. Il prend en compte de très nombreux facteurs comme l'espérance de vie, le produit intérieur brut par habitant, l'éducation, la santé, etc. . . Cette figure montre clairement le lien entre le niveau de développement d'un pays et l'usage des technologies numériques. Une langue parlée dans une région du monde pauvre ou en voie de développement, aura en général peu de diffusion sur le Web. Le contraire n'est pas forcément vrai, une langue d'une zone développée ne sera pas nécessairement une langue présente sur Internet. Le luxembourgeois par exemple dispose de très peu de visibilité sur le Web, non pas à cause du niveau de vie moyen des luxembourgeois, qui est plutôt élevé (16^e rang IDH), mais plutôt à cause de l'utilisation d'autres langues de communication dans ce pays, qui sont l'anglais, l'allemand et le français.

L'Internet est la principale source de collecte de textes et d'audio pour constituer des corpus de taille importante. Nous venons de voir que le nombre de locuteurs et le niveau

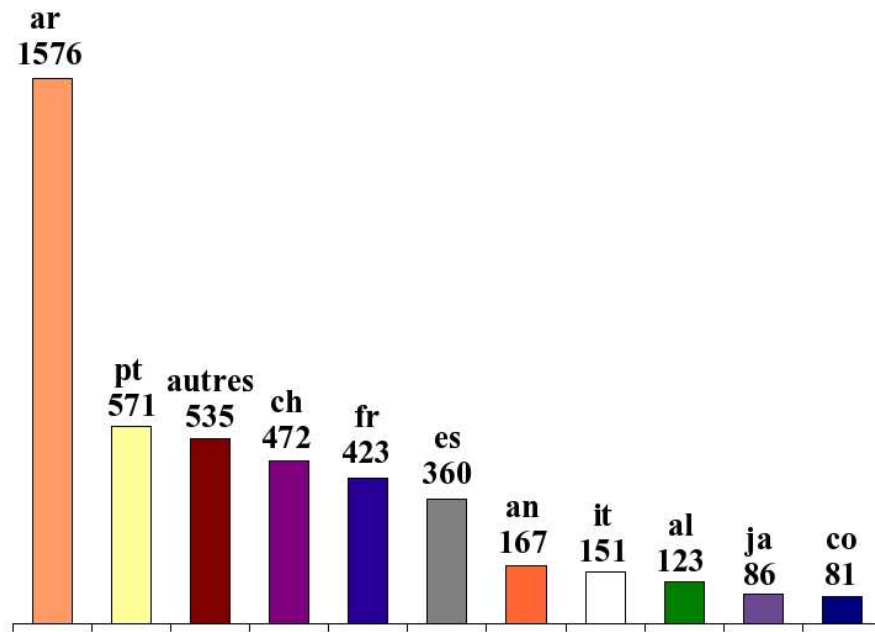


FIG. 1.6 – Évolution du nombre d'utilisateurs d'Internet par langue en pourcentage, entre 2000 et 2007. Cette étude date du 30 novembre 2007. Source : internetworldstats.com

de développement d'un pays ne sont pas des indicateurs fiables de l'importance de l'utilisation d'une langue sur le Web, et donc de la facilité de collecter des données. Essayons de préciser la définition d'une langue peu dotée.

Dans la thèse « Reconnaissance automatique de la parole pour des langues peu dotées » soutenue en novembre 2006 [Le, 2006], les **langues bien dotées** qui sont les quelques langues qui possèdent des ressources en quantité importante, sont opposées aux **langues peu dotées** qui disposent de peu de ressources linguistiques servant à élaborer les systèmes de reconnaissance. Nous utiliserons la même terminologie, en mettant l'accent sur le facteur quantité de textes qui nous semble être le facteur le plus limitant.

Ressources linguistiques pour la reconnaissance de la parole

Les trois types de données nécessaires à l'élaboration d'un système de reconnaissance de la parole actuel sont de grands corpus de textes (typiquement entre quelques dizaines et quelques centaines de millions de mots), un corpus audio de parole transcrite (typiquement entre quelques dizaines et centaines d'heures), ainsi qu'un lexique de mots donnés avec leur prononciation avec des variantes éventuelles.

Des corpus de données de projets de recherche peuvent éventuellement être récupérés.

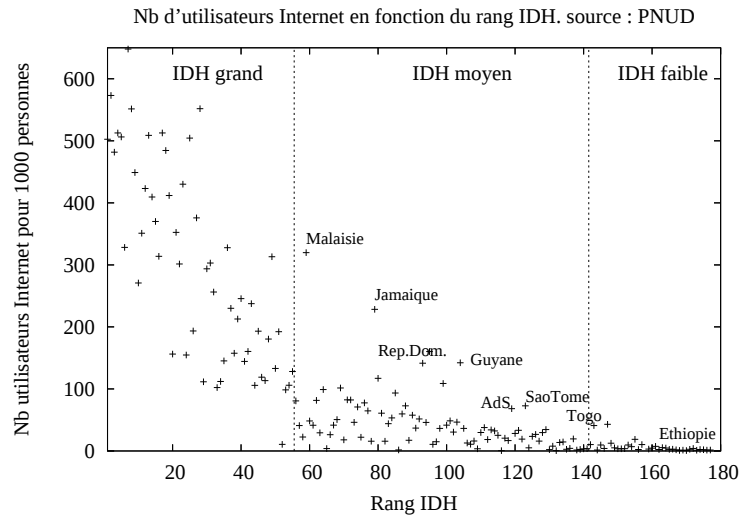


FIG. 1.7 – Nombre d'utilisateurs d'Internet sur des échantillons de 1000 personnes pour les pays classés selon l'Indice de Développement Humain (IDH). Source : UDNP

Nous avons mentionné dans l'introduction quelques uns des projets qui visent à réduire les coûts et le temps nécessaires à la conception de technologies de reconnaissance, et/ou de synthèse vocale pour des langues peu dotées, comme le projet « SPICE » par exemple [Schultz *et al.*, 2007], qui signifie « Speech Processing : Interactive Creation and Evaluation toolkit »⁵. À l'aide d'une interface Web interactive qui ne nécessite pas d'entraînement particulier pour un nouvel utilisateur, les ressources nécessaires au développement de ces technologies peuvent être intégrées facilement, et les performances de nouveaux systèmes évaluées de manière itérative. Les premières langues pour lesquelles SPICE a été utilisé sont l'afrikaans [Engelbrecht & Schultz, 2005], le bulgare [Mircheva, 2006] et le vietnamien [Le *et al.*, 2006]. En 2007, un cours de six semaines autour du projet a été organisé avec dix étudiants de langue native différente, dans le but d'utiliser l'interface pour chaque langue. Des langues aussi diverses que le bulgare, l'anglais, le français, l'allemand, l'hindi, le konkani (qui est l'une des 22 langues officielles parlées en Inde), le mandarin, le telugu (également une langue indienne officielle), le turc et le vietnamien [Schultz *et al.*, 2007]. Parmi les projets de base de données textes et parole multilingues plus anciens, figure par exemple « GlobalPhone » [Schultz, 2002], qui réunit des corpus pour 19 langues : arabe, mandarin, cantonais, allemand, français, japonais, coréen, croate, portugais, russe, espagnol, suédois, tamoul, tchèque, turc, thai, créole, polonais et bulgare. En ce qui concerne les corpus de parole de ce projet, il s'agit uniquement de parole lue, ce qui en limite un peu l'intérêt.

La langue étudiée peut également avoir fait l'objet d'une collecte de données par un

⁵SPICE : <http://cmuspice.org>

organisme de distribution de données linguistiques. Il en existe deux principaux : le consortium américain de données linguistiques LDC⁶, qui réunit des laboratoires publics, gouvernementaux et privés, et l'agence européenne de distribution de ressources linguistiques ELDA⁷. LDC propose un catalogue de plusieurs centaines de corpus de données de différents types : corpus de parole lue transcrite, de parole spontanée (conversations téléphoniques en particulier, CTS pour Conversationnal Telephone Speech), de parole d'émissions radio-télévision diffusion (BN pour Broadcast News), mais aussi des corpus de textes, principalement des textes provenant de journaux d'information (Newswire) ainsi que des lexiques et des dictionnaires de prononciations. De même, ELDA vend des corpus destinés à la recherche et à l'industrie. Les langues concernées sont très variées, allant des langues bien dotées (anglais, chinois, français, espagnol, etc...), à des langues peu dotées (langues des pays de l'Europe de l'est comme par exemple le slovène, le tchèque, l'estonien mais aussi des dialectes arabes, la langue turque, etc...).

Une collecte de corpus de textes et de parole peut également être réalisée sur le Web, à partir de sites d'émissions d'information par exemple. Des sites multilingues diffusent des programmes dans de nombreuses langues. Citons par exemple Radio France Internationale⁸, Deutsche Welle⁹, la British Broadcasting Corporation¹⁰, Voice of America¹¹, qui diffusent respectivement en 20, 30, 33 et 44 langues. Des problèmes de droit peuvent se poser, il est recommandé de demander l'autorisation avant de collecter les sites en question.

En ce qui concerne les textes, il existe des « robots », qui « aspirent » des sites Web entiers. Plusieurs robots, ou « Web crawlers » en anglais, dédiés à la collecte de textes pour les langues peu dotées existent déjà, on peut citer par exemple CorpusBuilder¹², An Crúbadán 2.0¹³, ainsi que la boîte à outils Clips-Text-Tk qui fournit également des outils de normalisation indépendants de la langue¹⁴. L'auteur du robot An Crúbadán a collecté des corpus de textes pour 416 langues peu dotées. Le tableau 1.4 donne quelques exemples de langues pour lesquelles un corpus de textes a été constitué avec ce robot. Le tableau donne le nombre de mots collectés, et le nombre de sites qui ont servi à constituer ces corpus. Ces chiffres sont tirés de la page sur le robot¹⁵. Des corpus, de langues aussi diverses que le bosniaque, le catalan ou le Gaélique ont été collectés, et totalisent respectivement 2,1, 2,7 et 98,1 millions de mots. Les plus petits corpus collectés totalisent un millier de mots environ, et sont pour la majorité des corpus de langues d'Afrique noire.

⁶The Linguistic Data Consortium : <http://www.ldc.upenn.edu>

⁷Evaluations and Language resources Distribution Agency <http://www.elda.org>

⁸RFI : <http://www.rfi.fr>

⁹DW : <http://www2.dw-world.de>

¹⁰BBC : <http://www.bbc.co.uk/>

¹¹VOA : <http://www.voanews.com>

¹²CorpusBuilder : <http://www.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/corpusbuilder>

¹³An Crúbadán : <http://borel.slu.edu/crubadan/stadas.html>

¹⁴Clips-Text-Tk : <http://www-clips.imag.fr/geod/User/brigitte.big/Logiciels.html>

¹⁵<http://borel.slu.edu/crubadan/stadas.html>

<i>Langue</i>	<i>Pays</i>	<i># Sites</i>	<i># mots collectés</i>
<i>gaélique</i>	Irlande	140k	98,1M
<i>bosniaque</i>	Bosnie Herzégovine	112	2,1M
<i>catalan</i>	Espagne	400	2,7M
<i>créole haïtien</i>	Haïti	790	2,5M
<i>baoulé</i>	Côte d'Ivoire	21	21k

TAB. 1.4 – *Quelques exemples de langues parmi les 416 langues pour lesquelles le robot An Crubádan a été utilisé pour collecter des textes sur le Web. Le pays où ces langues sont parlées sont indiqués, ainsi que le nombre de sites visités par le robot, et le nombre total de mots qui ont pu être collectés (M pour millions et k pour milliers).*

Source : <http://borel.slu.edu/crubadan/stadas.html>

Enfin il faut souligner l'importance du lexique de prononciations, qui fait le lien entre les modèles acoustiques et syntaxiques. Des informations linguistiques générales, des liens vers des sites de ressources, des radios-télévisions en ligne, des inventaires de phonèmes de nombreuses langues sont fournis par quelques sites, qui sont des aides précieuses dans l'élaboration d'un lexique de prononciations. Citons par exemple Omniglot¹⁶, Ethnologue Languages of the World¹⁷ ainsi que Wikipedia¹⁸.

Problèmes posés par le manque de données

Les trois types ressources présentés ci-dessus peuvent poser problème s'ils font défaut. Néanmoins, nous pensons que le manque de textes est le point le plus problématique, dans la mesure où il n'est pas possible de remédier, de quelque façon que ce soit, à l'absence de textes disponibles, due en particulier à un très petit nombre de sites Internet dans la langue étudiée. Les systèmes de transcription automatique sont pour la plupart des systèmes à vocabulaire fermé, c'est-à-dire que seuls les mots du lexique de prononciations peuvent être reconnus. Ainsi le manque de textes fait que les taux de mots hors-vocabulaire peuvent être très élevés, typiquement au dessus de 5%. D'autre part, les modèles de langage sont estimés sur très peu d'occurrences des différents n-grammes, et sont pour cette raison peu fiables.

La plupart des études sur le portage rapide de systèmes de reconnaissance à de nouvelles langues, ou sur la reconnaissance des langues peu dotées, se sont intéressées à la modélisation acoustique. Dans le projet sus-mentionné SPICE [Schultz *et al.*, 2007], des modèles multilingues sont utilisés pour initialiser les modèles acoustiques (technique dite de « bootstrap », présentée dans la section 1.4), et sont entraînés de manière itérative pour devenir dépendants de la langue cible. Dans [Le, 2006], des mesures de proxi-

¹⁶Omniglot : <http://www.omniglot.com>

¹⁷Ethnologue : <http://www.ethnologue.com>

¹⁸Wikipedia : <http://fr.wikipedia.org>

mité entre modèles acoustiques de phones sont proposées pour sélectionner les meilleurs modèles d’initialisation multilingues. Des efforts sont également souvent concentrés sur le développement d’outils destinés à collecter des données pour petit à petit rendre les langues peu dotées un peu plus dotées.

En ce qui concerne la création d’un lexique de prononciation, ou lexique de reconnaissance, l’approche la plus couramment utilisée lorsque peu de connaissances linguistiques sont accessibles sur la langue étudiée pour générer des prononciations, est d’associer un phone à chaque graphème. Cette approche est appelée modélisation acoustique à base de graphèmes. Elle a le double avantage de permettre la génération d’un lexique très simplement et très rapidement. Cette méthode a été étudiée pour différentes langues : arabe [Abdou, 2004], russe [Stücker & Schultz, 2004], vietnamien et khmer [Le, 2006], mais également pour des langues bien dotées : allemand, anglais, espagnol [Killer *et al.*, 2003], allemand, anglais, italien, néerlandais [Kanthak & Ney, 2002]. Dans [Le, 2006] par exemple, la langue Khmer a nécessité une romanisation de son système d’écriture. Aucun lexique de prononciation n’était disponible, et c’est donc un lexique basé sur les graphèmes qui a été utilisé. Pour cette langue, un taux d’erreurs de mots de 20,0% a été obtenu sur un corpus de parole lue. Toujours dans [Le, 2006], une comparaison de performances entre des systèmes qui utilisent une représentation en phonèmes et une représentation en graphèmes pour les modèles acoustiques, a été réalisée sur le vietnamien. Un taux d’erreurs de syllabes inférieur de 10% relatifs seulement a été obtenu par un système avec un lexique de prononciations basé sur les graphèmes, par rapport à un système basé sur les graphèmes, avec un taux d’erreurs de syllabes de référence de 47,8% sur de la parole d’émissions radio-télévision (Broadcast News). Les différences de performances entre des systèmes basés sur les graphèmes et les systèmes basés sur les phonèmes dépendent bien sûr de la relation qui peut être très directe pour certaines langues, comme le turc par exemple, ou plus complexe comme pour l’anglais. Dans [Kanthak & Ney, 2002], l’utilisation des graphèmes provoque une augmentation relative de seulement 2% du taux d’erreurs de mots par rapport à des modèles acoustiques de phonèmes, pour les trois langues bien dotées allemand, italien, néerlandais. En revanche pour l’anglais, l’augmentation du WER atteint 20% relatifs. Néanmoins, un lexique fondé sur les graphèmes ne comporte pas de variantes de prononciation. Nous proposerons dans le prochain chapitre une méthode pour générer des variantes à partir d’alignements phono-tactiques.

Influence de la quantité de données sur les performances

Dans la littérature, on trouve peu d’études sur l’influence de la quantité de données audio et/ou textes sur les performances des systèmes de reconnaissance. Dans [Lamel & Adda, 2000], l’usage de sous-titres automatiques, appelés « Closed Captions », ou de transcriptions détaillées, a été comparé pour évaluer l’entraînement des modèles acoustiques (MA). La langue concernée était l’anglais américain. L’entraînement des MA avec des closed captions est appelé entraînement peu supervisé ou semi-supervisé, dans la mesure

où ces sous-titres ne sont pas l'exacte transcription de ce qui est dit, mais plutôt un résumé qui permet aux mal-entendants et aux sourds de comprendre ce qui est dit. Les performances obtenues avec l'apprentissage peu supervisé sont du même ordre que celles obtenues avec un apprentissage supervisé classique avec des transcriptions fines, sur une tâche de parole de type radio-télévision. Le modèle de langage utilisé dans cette étude était très bon puisqu'il fut entraîné sur plus d'un milliard de mots (790 millions de mots provenant de textes de journaux d'information et 240 millions de mots de transcriptions commerciales de radio-télévision).

Dans [Lamel & Adda, 2002], l'étude a été étendue en mesurant les taux d'erreurs obtenus avec des quantités de données d'apprentissage comprises entre 10 minutes et 200 heures de données non-transcrites, tout en réduisant considérablement la quantité de textes utilisés pour l'estimation des modèles de langage à une taille totale de 1,8 millions de mots. Une procédure itérative a permis d'améliorer la qualité des modèles acoustiques en dépit d'un taux d'erreurs initial très élevé. Ces études ont montré que des transcriptions fines manuelles ne sont pas obligatoires pour obtenir des modèles acoustiques performants.

La figure 1.8 représente le taux d'erreurs de mots (WER pour Word Error Rate) du tableau 4 de [Lamel & Adda, 2002], en fonction de la quantité de données audio non-transcrites manuellement. L'apprentissage acoustique semi-supervisé conduit à un taux d'erreurs de mots de 37,4% en utilisant 135h de données automatiquement transcrites, avec un modèle de langage entraîné sur 1,8 millions de mots, alors que le taux d'erreurs obtenu avec seulement 10 minutes de données manuellement transcrites qui ont servies d'initialisation aux modèles était de 65,3%. Avec les mêmes 123h de données pour entraîner de manière supervisée, le taux d'erreurs atteignait 28,8% avec le même modèle de langage. Selon [Moore, 2003], ces courbes montreraient qu'il y a une loi linéaire entre le logarithme du taux d'erreurs et la quantité de données d'entraînement. Il faut cependant remarquer qu'à partir d'une certaine quantité de données, il y a un palier à partir duquel ajouter plus de données n'apporte pas de gain supplémentaire.

Les études [Lamel & Adda, 2000] et [Lamel & Adda, 2002] ont été menées sur l'anglais américain, qui est la langue la plus étudiée en reconnaissance de la parole. Les approches et les résultats peuvent être différents pour les langues peu dotées en général, pour lesquelles très peu voire aucune expertise linguistique n'est accessible. Réduire la quantité de données de manière artificielle pour une langue bien dotée pour mesurer l'influence sur les performances d'un système n'est pas équivalent à élaborer un nouveau système, pour une langue peu dotée. En effet toute l'expertise acquise au fil des années d'expérience sur une langue bien dotée influence les choix de développement. Les problèmes rencontrés pour un tout nouveau système ne sont a priori plus nombreux : choix de la représentation des modèles acoustiques, normalisation des textes et transcriptions, création du lexique de prononciations entre autres. Dans le chapitre 2, les influences de la quantité de données audio et de données textes seront étudiées conjointement pour une langue peu dotée.

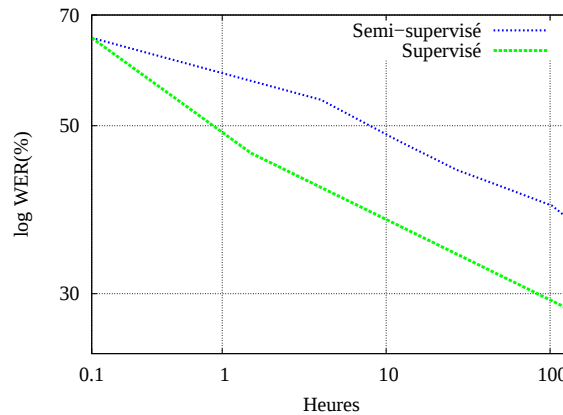


FIG. 1.8 – Taux d’erreurs de mots (%) en fonction de la quantité de données audio d’apprentissage (chiffres donnés dans [Lamel & Adda, 2002])

1.9 Conclusion

Nous avons présenté le principe de la reconnaissance de la parole par modèles statistiques, et décrit brièvement les différentes composantes d’un système standard : modèles acoustiques de phones, modèles de langage, lexique de prononciation, et présenté très succinctement le principe des décodeurs. Des taux d’erreurs indicatifs ont été donnés pour des systèmes de reconnaissance à l’état de l’art pour la langue la plus étudiée, l’anglais américain. En particulier pour de la parole provenant d’émissions de radio-télévision, le taux d’erreurs de mots est d’environ 10% sans aucune restriction sur les données. Pour arriver à de telles performances, de très grands corpus de données sont nécessaires pour estimer les paramètres des modèles. En fin de chapitre, nous avons précisé les problèmes posés par le manque de données d’apprentissage, et défini ce que sont les langues peu dotées vis-à-vis de la reconnaissance de la parole. Le manque de textes pour ces langues est à nos yeux le problème le plus central. Aux difficultés liées au manque de données s’ajoute en général le manque d’expertise et d’informations linguistiques. Comment élaborer un lexique de prononciations ? Quelles sont les performances de reconnaissance d’un système en fonction des quantités de données d’apprentissage pour une langue peu dotée ? Le prochain chapitre tentera de répondre à ces questions sur un cas d’étude.

Chapitre 2

Reconnaissance automatique de l'amharique

Ce chapitre a pour but de montrer quels ont été les problèmes pratiques que nous avons rencontré lors de l'élaboration d'un système de reconnaissance pour l'amharique : normalisation des textes et génération du lexique de prononciations en particulier.

Nous décrivons également une étude expérimentale sur l'influence comparée des quantités d'audio transcrit, de transcriptions et de textes utilisés pour l'apprentissage des modèles acoustiques et modèles de langage. Enfin, nous montrerons les résultats obtenus lors d'une première expérience de décompositions des unités lexicales, qui a été le point de départ des recherches sur la modélisation lexicale qui fera l'objet des chapitres suivants.

2.1 Présentation de la langue amharique

L'amharique est la langue officielle de la République Démocratique Fédérale d'Éthiopie. La carte 2.1 situe l'Éthiopie au sein du continent africain (source : Wikipedia).

L'amharique est parlé par environ 22 millions de locuteurs dont 17 millions comme langue maternelle, et 5 millions comme seconde langue [Appleyard, 1995]. Si l'amharique est la langue la plus parlée en Éthiopie, il existe cependant plus de 80 langues différentes et quelques 200 dialectes. La seconde langue la plus importante est l'oromo, parlée par plus de 17 millions de locuteurs.

Bien que faisant partie des langues sémitiques comme l'arabe et l'hébreu, l'amharique possède une écriture de gauche à droite, avec un syllabaire spécifique appelé « Fidel », terme qui signifie également « lettre, caractère ». Le Fidel est dérivé de la langue classique éthiopienne, le ge'ez. Il possède 34 symboles de base dont 85% représentent une séquence CV (C pour consonne, V pour voyelle), les autres symboles représentent une séquence

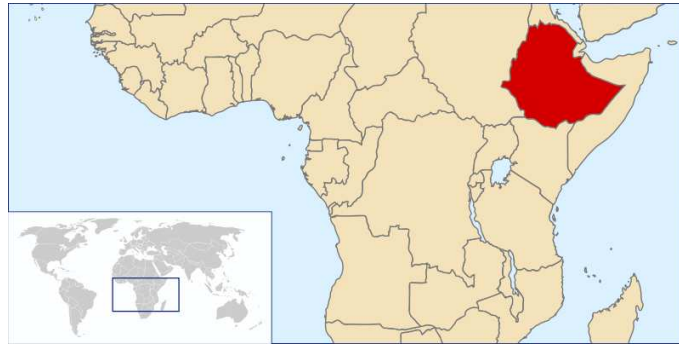


FIG. 2.1 – Situation géographique de l'Éthiopie (source : Wikipedia).

CwV où w est une semi-consonne. Un dernier symbole représente le son complexe [ts]. Cette langue possède sept voyelles au total, schwa inclus, appelées les sept ordres : [ɛ], [u], [i], [a], [e], [ə] et [o]. Les sept ordres sont indiqués à l'écrit par une modification du signe de base des consonnes. Il existe également des symboles redondants, qui représentent une même syllabe. Il y a donc au total plus de 240 symboles différents.

Il faut remarquer qu'il existe des problèmes de normalisation de l'orthographe amharique, très bien exposés dans [Yacob, 2003], où trois niveaux de langue sont distingués :

- l'amharique canonique qui est réservé aux érudits avec une orthographe unique par mot,
- l'amharique commun qui est celui des journaux, de la littérature avec de nombreuses formes homophones et des variantes orthographiques inter-individuelles,
- l'amharique quotidien pour lequel aucun jugement n'est porté sur l'orthographe des mots.

L'orthographe des mots amhariques utilisés au quotidien est très libre, le nombre de formes écrites différentes pour un même mot peut être très grand. L'exemple suivant montre la diversité des formes que nous avons rencontré dans les corpus de textes. Il s'agit de l'entité nommée « Francfort Germany », mais le phénomène est tout à fait général sur les mots amhariques. Comme nous le détaillerons ci-dessous, les caractères amhariques ont été transcodés à l'aide d'un jeu de lettres latines, et ici le « x » représente un schwa. Dans « Germany », le /a/ est remplacé par un /E/. « Francfort » est écrit soit en un mot, soit en deux mots, avec des confusions possibles dans les voyelles, ici entre /E/ et /o/.

fxranxkxfErxtx	JErxmani
fxranxkxforxtx	JErxmani
fxranxkx fErxtx	JErxmEni
fxranxkxfErxtx	JErxmEni

Quelques propriétés lexicales

La figure 2.2 montre le nombre de mots distincts du corpus audio en fonction de la taille des mots en phones (50,3k mots distincts). La longueur des mots la plus fréquente est de 10 phones soit 5 syllabes. Cette longueur relativement grande s'explique par l'agglutination d'affixes pour les articles, les démonstratifs, les marques de pluriel entre autres [Demisse & Imbert-Vier, 1996].

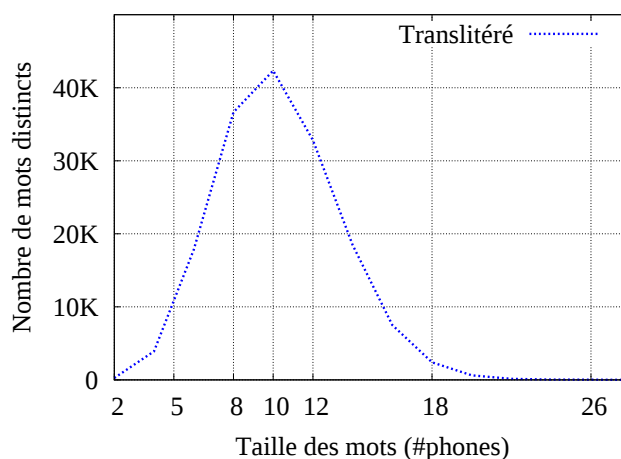


FIG. 2.2 – Distribution des lexèmes en fonction de leur taille en phones.

La figure 2.3 montre l'évolution de la taille moyenne des mots du corpus audio en fonction de leur rang de fréquence. Les mots les moins fréquents ont une longueur moyenne deux fois supérieure à celles des mots les plus fréquents.

2.2 La langue amharique, une langue peu dotée ?

Bien que l'Éthiopie soit un pays économiquement pauvre (l'Éthiopie est au 170e rang mondial selon l'indice de développement humain [The United Nations Development Program, 2006]), la production littéraire et culturelle en général a toujours été très importante. Les Éthiopiens sont fiers d'appartenir à un peuple très ancien, dans un pays qui est le berceau de l'humanité. L'amharique est une langue à laquelle ils sont très attachés.

La langue amharique et le traitement automatique

L'amharique fait l'objet d'études en traitement automatique, comme par exemple sur la normalisation de l'orthographe [Yacob, 2003], sur le classement thématique de textes

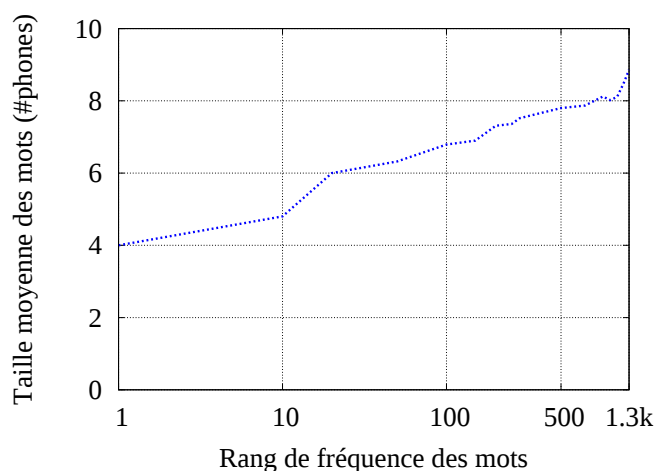


FIG. 2.3 – Longueur moyenne des mots (en nombre de phones) en fonction du rang de fréquence.

[Eyassu & Gambäck, 2005], sur les outils d'analyse morphologique [Fissaha & Haller, 2003]. Il existe un portail Web de type wiki sur le traitement automatique de l'amharique qui vise à regrouper les ressources, corpus et publications scientifiques¹. Par ailleurs, il existe plusieurs systèmes de translittération dont le plus connu est le système SERA (System for Ethiopic Representation in ASCII) qui sert à écrire en Fidel avec un clavier latin².

En ce qui concerne la transcription automatique de la parole, les études les plus récentes sont celles de Hussien Seid et Björn Gambäck ([Seid & Gambäck, 2005]), et de Solomon Teferra Abate et al. ([Abate *et al.*, 2005; Abate & Menzel, 2007]). Ces études décrivent la mise en place de systèmes de reconnaissance indépendants du locuteur. Comme il n'existe pas de corpus de parole disponible en amharique les auteurs ont construits des corpus de parole lue. Dans [Seid & Gambäck, 2005], les auteurs annoncent un taux d'erreurs sur les mots de 25,7% obtenu sur un corpus de test très petit, composé de deux phrases prononcées par dix locuteurs différents non-présents dans le corpus d'apprentissage. Dans cette étude, le vocabulaire est fermé, et contient moins de mille mots.

Dans [Abate *et al.*, 2005], la création d'un corpus phonétiquement équilibré de parole lue est décrite. Dans [Abate & Menzel, 2007], un système à vocabulaire fermé construit avec HTK [Young, 1993] est décrit. Les modèles acoustiques sont appris sur 20 heures de transcriptions de parole lue, et un modèle de langage bigramme est estimé sur moins d'un million de mots. Des taux d'erreurs de mots autour de 9% sont présentés.

¹<http://corpora.amharic.org>

²<http://www.abysiniacybergateway.net/fidel>

La langue amharique et les ressources numériques

Des études de traitement automatique de l'amharique, il ressort que les ressources numériques disponibles (corpus, lexique et outils même basiques comme des analyseurs ou des étiqueteurs morphologiques) sont très réduites [Asker *et al.*, 2007].

Les corpus audio décrits dans la littérature sont tous des corpus de parole lue. Remarquons que la parole lue est un style de parole qui n'est plus guère étudié pour les langues bien dotées. Sont préférés des styles de parole de données plus « réelles », comme la parole provenant d'émissions de radio et de télévision dite parole « broadcast news ». Les corpus de textes sont très limités en nombre de mots, le plus grand totalisant moins d'un million de mots [Abate & Menzel, 2007].

Il existe cependant quelques journaux et radios en ligne sur Internet. Par exemple deux radios diffusent des émissions d'information : la radio internationale allemande Deutsche Welle ³, qui diffuse en 30 langues dont l'amharique, et la radio éthiopienne « The voice of Ethiopian Medhin »⁴.

2.3 Les données audio et textes

Pour élaborer les systèmes de transcription utilisés au cours de cette thèse, deux types de données ont été utilisés : un corpus d'émissions de radio transcrites et un corpus de textes issus de sites Web de journaux en ligne. Le corpus audio contient 37 heures d'émissions de type journal provenant de radio Deutsche Welle (25h) et radio Medhin (12h) enregistrées de janvier 2003 à janvier 2004. Ces données ont été transcrites manuellement par des locuteurs éthiopiens. Pour tester et développer le système de transcription automatique, deux heures de données audio transcrites ont été sélectionnées au sein du corpus. Il s'agit des fichiers audio parmi les plus récents du corpus, les thèmes abordés dans ces données pouvant être nouveaux par rapport à ceux des données d'apprentissage. Ce corpus de devtest couvre les mois de décembre 2003 et janvier 2004. Le corpus de développement est très proche en date du corpus d'apprentissage, et il est très probable que certains locuteurs du corpus de développement soient également présents dans le corpus d'apprentissage. Nous n'avons donc pas de réel corpus de test. Le tableau 2.1 résume les caractéristiques des données audio : le nombre d'heures par source, le nombre de locuteurs et le nombre de mots pour le corpus d'apprentissage et pour le corpus de développement. Nous avons retenu un plus grand nombre d'heures provenant de Deutsche Welle que de radio Medhin, car les émissions de cette radio montrent une plus grande diversité de locuteurs.

Les données textuelles autres que les transcriptions manuelles proviennent de 3 sources : Ethiozena (archives de 1988 à 1996 et textes récents), Deutsche Welle (textes récents) et Ethiopian Reporter (textes récents). Au total pour ces trois sources, nous disposons

³www2.dw-world.de

⁴www.medhininfo.com

Source	Train	devtest
Deutsche welle	24h06	1h20
radio Medhin	11h08	0h37
# locuteurs	200	15
# mots	232,6k	14,1k

TAB. 2.1 – Nombre d’heures, de locuteurs et de mots pour les deux sources audio.

de 4,6 millions de mots. Les textes des transcriptions du corpus audio sont également utilisées et totalisent 246,7k mots.

2.4 L’étape de normalisation

Le code ASCII 7 bits permet d’utiliser 128 caractères différents. Pour représenter les 240 symboles amhariques il faudrait au minimum un alphabet à 8 bits. En réalité si l’on considère tous les symboles amhariques – avec des variantes homophones et des symboles pour les nombres – il y a plus de 256 symboles différents et il faudrait un alphabet à 9 bits). Le code de translittération SERA mentionné ci-dessus associe plusieurs lettres latines pour certains symboles du Fidel pour pouvoir utiliser uniquement des caractères ASCII. Pour des raisons pratiques, les phones ne sont habituellement représentés que par un seul caractère. SERA n’était donc pas utilisable pour définir un jeu de phones amharique.

Nous avons choisi un jeu de caractères latins interne au LIMSI qui reprend des symboles de phones usuels pour des systèmes de reconnaissance d’autres langues. Le jeu de caractères latins a été élaboré à partir de quelques ouvrages d’apprentissage de l’amharique principalement en langue anglaise [Appleyard, 1995; Isenberg, 1842] mais aussi en français [Demisse & Imbert-Vier, 1996]. Le jeu comporte 33 lettres (7 voyelles et 26 consonnes). Le lien entre graphèmes et phonèmes est relativement direct en amharique : nous avons donc pu couvrir tous les symboles amharique avec le même jeu de caractères qui sert pour désigner les phones (modèles acoustiques). Les 33 caractères représentent donc à la fois les graphèmes et les phonèmes. La table de correspondance des caractères est donnée dans l’annexe 6.5.

Le tableau 2.2 donne l’exemple des sept ordres pour la syllabe **ላ** transcrite par /lE/. Le noyau vocalique du sixième ordre correspond à un schwa, noté x.

Dans la section 3.2, nous avons défini la différence entre transcription et translittération. Dans cet exemple, la conversion des symboles ge’ez à l’aide d’un jeu de caractères latins s’apparente à une transcription dans la mesure où des graphèmes ge’ez homophones ont été associés aux mêmes suites de caractères latins. Par exemple, il existe quatre graphèmes différents qui représentent le même son [h]. Par souci d’économie et de lisibilité, ces

Symboles ge'ez	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
Symboles transcrits	lE	lu	li	la	le	lx	lo

TAB. 2.2 – *Exemple des différents ordres ou noyaux vocaliques associés à une consonne en symboles ge'ez et transcrits. Le sixième ordre transcrit par la lettre x correspond à un schwa.*

graphèmes ont été transcrits en la seule lettre h. Il en est de même pour tous les graphèmes homophones en général, ainsi n'est-il pas possible de revenir de manière non-ambiguë à la forme initiale ge'ez d'un mot transcrit en caractères latins. Pour cette raison, cette normalisation est plus une transcription qu'une translittération. Les étapes suivantes de normalisation sont les étapes habituelles de transformation des dates et des nombres en mots entiers. Nous avons pu établir la liste et le format des dates et des nombres amhariques à l'aide des trois méthodes et grammaires d'amharique [Demisse & Imbert-Vier, 1996; Appleyard, 1995; Isenberg, 1842].

2.5 Le lexique de prononciation

Un lexique de 133k mots a été sélectionné à partir de tous les mots distincts des transcriptions (50k mots) et des mots des textes Web apparaissant au moins trois fois. Ne pas prendre les mots n'apparaissant qu'une ou deux fois dans le corpus collecté sur le Web permet de filtrer les hapax et la plupart des mots mal écrits. Le taux de mots hors vocabulaire ou taux d'OOV (Out-Of-Vocabulary) du corpus de développement avec ce lexique est de 7,0% ce qui est très élevé par rapport aux langues bien représentées qui ont des taux d'OOV autour de 1 ou 2% avec des lexiques de 65k mots pour l'anglais et le français par exemple.

Le tableau 6.13 donne les caractères latins que nous avons choisis pour normaliser les symboles ge'ez. Les tableaux 2.3 et 2.4 fournissent la liste des consonnes et des voyelles amhariques respectivement avec les symboles de l'API, en indiquant le lieu et la manière d'articulation.

Les sons qui n'ont pas d'équivalent en français ou en anglais sont les consonnes dites glottalisées ou éjectives : [p'], [d'], [tʃ'] et [s']. Chacune de ces consonnes glottalisées a une consonne correspondante non-glottalisée. Beaucoup de mots diffèrent uniquement d'une consonne glottalisée ou non [Appleyard, 1995].

Comme les caractères amhariques sont transcrits avec un jeu de caractères représentant les phones eux-mêmes, un premier dictionnaire de prononciation constitué de la simple liste des mots a été utilisé pour réaliser les premiers alignements. Ces alignements ont permis de constater que les schwas ne sont pas toujours prononcés, et un deuxième dictionnaire avec tous les schwas optionnels a été utilisé. Globalement, plus de 60% des

	<i>Lieu d'articulation</i>				
	<i>Labial</i>	<i>Dental</i>	<i>Palatal</i>	<i>Vélaire</i>	<i>Glottal</i>
<i>Plosives</i>	[p], [b], [p']	[t], [d], [d']	[tʃ], [dʒ], [tʃ']	[k], [g], [q], [k ^w], [g ^w], [q ^w]	[ʔ]
<i>Fricatives</i>	[f]	[s], [z], [s']	[ʃ], [ʒ]		[h]
<i>Nasales</i>	[m]	[n]	[ɲ]		
<i>Liquides</i>		[l], [r]			
<i>Glides</i>	[w]			[j]	

TAB. 2.3 – Inventaire des phonèmes consonantiques de l'amharique (symboles API).

	<i>Antérieures</i>	<i>Centrales</i>	<i>Postérieures</i>
<i>Fermées</i>	[i]	[ɨ]	[u]
<i>Semi-fermées</i>	[e]	[ə]	[o]
<i>Ouvertes</i>		[a]	

TAB. 2.4 – Inventaire des voyelles de l'amharique (symboles API).

schwas sont élidés lorsqu'ils sont optionnels.

Le tableau 2.5 donne l'exemple de trois mots de ce lexique, avec leur rang de fréquence et leur nombre d'occurrences au sein des transcriptions manuelles. Le schwa est mis entre accolades pour signifier son caractère optionnel dans le lexique de prononciations. Sur les 3k occurrences de « nEwx » qui est le mot le plus fréquent des transcriptions, 2,5k mots ont été alignés avec la prononciation sans schwa.

<i>Lexème</i>	<i>Prononciation</i>	<i>rang</i>	<i>#occ.</i>
nEwx	nEw{x}	1	3044
mEto	mEto	7	803
jEdimokxras	jEdimok{x}rasi	236	47

TAB. 2.5 – Exemples de mots du lexique avec leur prononciation, rang de fréquence et nombre d'occurrences.

Le lexique de prononciations décrit ci-dessus est similaire à un lexique fondé sur les graphèmes. La différence réside dans la transcription des caractères ge'ez en alphabet latin. Un véritable lexique fondé sur les graphèmes aurait associé directement un phone à chaque symbole ge'ez. Mise à part l'étape de transcription qui a nécessité des connaissances linguistiques, les prononciations principales du lexique amharique sont identiques aux entrées lexicales : en ce sens, l'approche est une approche purement graphémique.

2.6 Génération de prononciations à l'aide d'alignements phonotactiques

Nous avons élaboré une méthodologie qui permet d'identifier et de valider des variantes de prononciation sans connaissance linguistique a priori. Cette approche est fondée sur les corpus d'apprentissage des modèles acoustiques. L'idée consiste à faire des alignements au niveau phonémique avec un lexique qui autorise la substitution de chaque phone par n'importe quel autre phone, et de sélectionner a posteriori les substitutions les plus fréquentes.

Un système d'alignement au niveau des graphèmes (pour l'amharique, le niveau graphémique correspond au niveau syllabique) est utilisé avec un lexique qui comporte toutes les prononciations possibles pour chaque syllabe. En d'autres termes, chaque syllabe peut se voir substituée par n'importe quelle autre syllabe. En pratique, deux alignements ont été réalisés, l'un autorisant les substitutions et élisions uniquement sur les voyelles et l'autre uniquement sur les consonnes dans le but d'éviter d'avoir trop de variantes simultanément.

Pour la première étape d'identification de variantes potentielles, des modèles indépendants du contexte, ou monophones, ont été utilisés avec une représentation lexicale syllabotactique. Pour la seconde étape de validation ou de sélection des variantes les plus fréquentes, la représentation lexicale est fondée sur les mots entiers, et plusieurs jeux de modèles acoustiques ont été testés, un jeu de monophones et plusieurs jeux de modèles dépendants du contexte.

Identification des variantes

Pour cette première étape, le lexique est simplement la liste de toutes les syllabes de l'amharique. En plus de la prononciation principale, identique à la forme orthographique, toutes les substitutions ainsi que l'élision des voyelles ont été ajoutées. Il a été procédé de même sur un deuxième lexique identique, avec toutes les substitutions possibles des consonnes, les voyelles étant gardées inchangées.

Le tableau 2.6 donne un exemple de représentation syllabique pour un mot amharique.

<i>Mot</i>	<i>Forme syllabotactique</i>
ቦደሞክረሲ	ቦ_ _ደ_ _ሞ_ _ከ_ _ረ_ _ሲ_
bEdemokxrasi	bE_ _de_ _mo_ _kx_ _ra_ _si

TAB. 2.6 – Exemple d'entrée lexicale avec sa forme syllabotactique associée. Le signe underscore sert à conserver l'information concernant la position des syllabes dans le mot.

Deux entrées pour des syllabes avec un 'd' initial sont données dans le tableau 2.7 avec

les prononciations associées, les prononciations principales sont en gras.

_dE	də	du	do	di	de	da	dɛ	d
dE	də	du	do	di	de	da	dɛ	d
dE_	də	du	do	di	de	da	dɛ	d
_da	də	du	do	di	de	da	dɛ	d
da	də	du	do	di	de	da	dɛ	d
da_	də	du	do	di	de	da	dɛ	d

TAB. 2.7 – *Extrait du lexique syllabique. Les prononciations « principales » sont indiquées en gras.*

Le tableau 2.8 est un exemple de matrice de confusion (en %) dont les colonnes donnent les formes graphémiques de la transcription de référence et les lignes donnent les formes phonémiques des voyelles sélectionnées lors de l'alignement forcé. La première ligne (#Occurrences) donne le nombre d'occurrences des voyelles dans le corpus et la dernière ligne donne le pourcentage d'élision. Il y a bien un peu plus de 60% de schwas qui sont élidés. La voyelle /E/ semble instable avec moins de 50% des occurrences qui sont alignées comme [E]. La substitution la plus fréquente est celle de /E/ en [a] (10,6%), et le taux d'élision est également élevé, supérieur à 20%.

Trois paires de voyelles avec des taux de substitution élevés sont indiquées en gras dans le tableau : /E/–[a] ; /e/–[i] ; /o/–[u]. Ces paires qui présentent une confusion importante rappelle le jeu de voyelles standard des langues sémitiques /a, u, i/ comme celui de l'arabe.

# Occurrences	Graphèmes						
	E	a	e	i	o	u	x
	290k	156k	22k	51k	41k	40k	363k
ɛ	47,4	7,1	6,1	1,5	5,7	1,9	4,8
a	10,6	78,0	1,4	1,0	2,0	0,6	1,5
e	6,7	0,7	64,0	11,0	1,5	0,9	2,7
i	1,9	0,3	11,2	53,8	1,4	1,8	4,2
o	4,4	1,0	0,9	0,8	67,1	14,8	1,8
u	1,4	0,2	0,6	1,2	10,7	57,1	3,5
ə	5,7	0,5	2,3	5,1	3,7	5,7	20,0
Élision	21,8	12,2	13,5	25,5	8,1	17,2	61,4

TAB. 2.8 – *Matrice de confusion entre voyelles. Les modèles acoustiques utilisés ici sont des monophones.*

La matrice 2.8 donne les pourcentages de substitution et élision globaux pour les voyelles alignées automatiquement. Pour avoir des estimations plus précises sur les variantes de prononciation des voyelles, les confusions entre voyelles ont été déterminées en fonction de l'élément consonantique qui précède le noyau vocalique (de type C ou Cw).

Le tableau 2.9 donne les pourcentages de substitution/élision pour les syllabes commençant par un 'b'. Les colonnes correspondent aux formes graphémiques avec le nombre d'occurrences dans la première ligne et les différentes formes phonémiques dans les lignes suivantes. La première colonne montre que 55% des syllabes 'bE' sont alignées avec la prononciation principale [bE], 16,9% avec la variante [b] (élision de la voyelle) et 10,6% avec la variante [bo]. Ces résultats diffèrent des tendances globales indiquées dans le tableau 2.8 où la voyelle de substitution la plus fréquente pour /ε/ est [a]. Pour la syllabe 'be', la voyelle est plus stable (76,3% de [e], 2,8% d'élisions), et la confusion la plus fréquente concerne la voyelle /i/, ce qui correspond à la tendance générale pour cette voyelle. Les deux autres voyelles (a et o) sont stables avec des taux d'élision et de substitution autour de 20% globalement.

#Occ	Graphèmes						
	bE	ba	be	bi	bo	bu	bx
	30,5k	12,3k	2,2k	2,7k	1,3k	2,5k	12,8k
ε	54,8	3,0	4,4	1,1	4,0	1,1	7,0
a	9,1	87,4	1,0	1,4	1,2	0,5	1,6
e	2,1	0,2	76,4	11,1	0,6	0,3	2,5
i	0,6	0,3	11,6	53,8	0,4	0,8	4,8
o	10,6	2,3	1,5	1,1	79,7	21,1	5,9
u	2,2	0,1	0,6	1,7	10,3	61,9	8,9
ə	3,7	0,1	1,7	2,7	0,6	1,6	13,8
Élision	16,9	6,6	2,8	27,1	3,2	12,7	55,5

TAB. 2.9 – Matrice de confusion pour les syllabes commençant par 'b'. Les modèles acoustiques utilisés ici sont des monophones.

Validation sur un lexique de mots entiers

À partir des variantes identifiées au niveau syllabique, des expériences au niveau lexical ont été menées pour évaluer le besoin et la pertinence des variantes.

Des règles de substitution et d'élision ont été dérivées pour chaque syllabe à l'aide de matrices de confusion similaires à la matrice 2.9. Une seule variante de prononciation par syllabe a été retenue, en l'occurrence la plus fréquente. Par exemple pour la syllabe 'bE' l'élision de la voyelle est la variante la plus fréquente. Pour la syllabe 'be', c'est la voyelle [i] qui a été retenue. Avec la prononciation principale et une variante par syllabe, chaque entrée lexicale se voit attribuer 2^N séquences phonémiques où N est le nombre de syllabes du mot. Le choix d'inclure une variante par syllabe assure de fournir des variantes pour les mots courts (les dix mots les plus fréquents sont dissyllabiques) tout en évitant de générer trop de variantes pour les mots longs. Généralement les mots fréquents sont sujets à des variations de prononciation (en particulier des réductions), quelles que soient les langues.

Le tableau 2.10 donne des exemples d'entrées lexicales. Dans les formes phonémiques, les voyelles entre accolades sont optionnelles et les voyelles entre crochets sont interchangeables. Ainsi, les deux voyelles du mot nEwx sont optionnelles, ce qui peut donner la transcription [nw]. La voyelle /o/ du mot mEto peut être substituée par un [u]. Enfin la voyelle /e/ dans jEdemokxrasî peut être prononcée [i].

<i>Entrée lexicale</i>	<i>Forme phonémique</i>
nEwx	n{E}w{x}
mEto	m{E}t[ou]
jEdemokxrasî	j{E}d[ei]m[ou]k{x}r{a}s{i}

TAB. 2.10 – *Extrait du lexique de prononciations.*

Pour mesurer l'intérêt des variantes de prononciations, le taux « *variant2+* » a été proposé dans [Adda-Decker & Lamel, 1999]. Ce taux est le pourcentage de mots qui n'ont pas été alignés avec la prononciation principale, qui est la prononciation la plus fréquente. Pour l'amharique, nous avons choisi la prononciation principale comme la prononciation identique à l'entrée orthographique, pour laquelle tous les phones sont requis.

Des alignements ont été réalisés avec différents jeux de modèles acoustiques : 36 modèles indépendants du contexte et des modèles triphones dépendants du contexte : 100, 300, 1k et le maximum, soit 7,2k modèles. Le corpus audio utilisé pour mesurer les taux *variant2+* est le corpus audio d'apprentissage des modèles acoustiques ce qui peut entraîner une sous-utilisation des variantes de prononciation puisque les modèles acoustiques peuvent éventuellement modéliser implicitement ces variantes.

Le tableau 2.11 donne les résultats obtenus avec un jeu de 100 modèles dépendants du contexte pour deux mots dissyllabiques. Pour chaque mot, sont donnés le rang de fréquence, le nombre d'occurrences dans le corpus (#Occurrences), le nombre d'occurrences alignées (#Alignées), le taux *variant2+* et différentes formes phonémiques avec leurs pourcentages respectifs d'utilisation dans les alignements. Le mot le plus fréquent du corpus, nEwx, est prononcé avec tous ses phones dans seulement 10% des cas environ. La réalisation la plus fréquente est celle où le schwa est élide (81,5% des occurrences). La dernière forme phonémique avec les deux voyelles élidées n'est jamais utilisée. Le deuxième exemple a un comportement différent. Presque la moitié des occurrences est alignée avec la forme principale. Les variantes pour ce mot concernent l'élision potentielle de /E/ et la substitution de /o/ par [u]. La deuxième forme phonémique la plus fréquente est [mɛtu] (36,1%). Les formes avec élision de la première voyelle sont peu fréquentes (6,4% et 7,0%).

Ces deux exemples reflètent les tendances globales des alignements : les mots avec des schwas sont principalement alignés avec des formes réduites (avec élision des schwas). Les mots sans schwa sont alignés avec moins de variantes. Le taux *variant2+* des mots qui comportent des schwas est de 86,9% contre 51,3% pour les mots qui n'en comportent

<i>Mot</i>	<i>Rang</i>	<i>#Occurrences</i>	<i>#Alignées</i>	<i>Variant2+</i>	<i>Phones</i>	<i>%</i>
nEwx	1	3044	2964	90,3%	nɛwə	9,7
					nwə	8,7
					nɛw	81,5
					nw	0,0
mEto	7	803	783	49,5%	mɛto	50,4
					mto	6,4
					mɛtu	36,1
					mtu	7,0

TAB. 2.11 – *Taux Variant2+ et pourcentages d'alignement pour les variantes de prononciation de deux mots fréquents, avec un jeu de 100 modèles acoustiques dépendants du contexte.*

pas.

La figure 2.4 montre le taux variant2+ en fonction du rang de fréquence des mots pour différents jeux de modèles acoustiques : le jeu de 36 monophones et plusieurs jeux de modèles dépendants du contexte (100, 300, 1k et 7,2k modèles). Quatre points par courbe sont donnés : pour des rangs de fréquence de 10, 100, 1k et 10k. Ces points sont des moyennes des taux variant2+ des mots dont le rang est centré sur ces valeurs. Le nombre de mots pris pour calculer ces moyennes dépend de ces valeurs de référence. Le tableau 2.12 donne les rangs des mots utilisés pour chaque moyenne, le nombre moyen d'occurrences associé (# Occ. Moy.), la longueur moyenne des mots et le pourcentage de mots qui comportent des schwas (%schwa).

<i>Rang</i>	<i>Rangs</i>	<i># Occ. Moy.</i>	<i>Taille</i>	<i>%schwa</i>
10	4-16	710	5,4	72%
100	74-126	220	6,9	80%
1k	899-1101	30	8,5	82%
10k	9499-10501	3	10,0	85%

TAB. 2.12 – *Intervalles utilisés pour moyennner les taux variant2+.*

Les courbes confirment bien le fait décrit dans [Adda-Decker & Lamel, 1999] que les modèles acoustiques dépendants du contexte modélisent implicitement les variations acoustiques. Plus le nombre de contextes modélisés est grand, moins le nombre de variantes utilisées pour aligner est grand. La courbe avec le jeu de 36 modèles indépendants du contexte présente les taux variant2+ les plus élevés et la courbe du jeu de 7,2k modèles les taux les plus faibles. Pour tous les jeux de modèles, les taux variant2+ augmentent avec le rang de fréquence. Cela peut s'expliquer par deux facteurs : la taille moyenne des mots augmente avec le rang et le nombre de mots qui comportent des schwas augmente également. La taille moyenne des mots double entre les mots les plus fréquents (rang

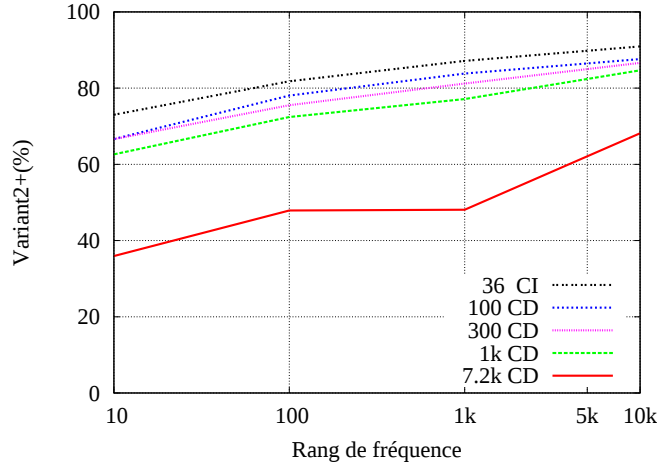


FIG. 2.4 – Taux Variant2+ en fonction du rang de fréquence des mots pour les jeux de modèles acoustiques suivants : 36, 100, 300, 1k et 7,2k modèles.

autour de 10) et les mots les moins fréquents (rang autour de 10k). 85% des mots les moins fréquents ont au moins un schwa et sont alignés principalement avec des formes réduites.

2.7 Modèles de phones amhariques

Les tableaux 2.3 et 2.4 donnent les 33 phones utilisés pour l'amharique, avec 26 consonnes et 7 voyelles. En plus des 33 phones, trois phones supplémentaires sont ajoutés, de manière habituelle, pour modéliser le silence, les respirations et les hésitations. D'autres études récentes décrivant des systèmes de reconnaissance pour l'amharique font état d'un jeu de phones plus important comprenant 38 phones comme par exemple dans [Abate *et al.*, 2005]. La différence vient du fait que nous modélisons les quatre consonnes vélarisées k^w , g^w , q^w et h^w par deux phones et non par un nouveau phone pour chacun. De même, le cluster 'ts' est modélisé par les deux phones [t] et [s].

Après avoir établi le jeu de phones, les modèles de Markov cachés (HMM) correspondants sont choisis à l'aide de modèles dits de « bootstrap » ou modèles « seed ». Ces modèles sont issus de langues déjà traitées. Pour l'amharique comme pour le turc qui fera l'objet du dernier chapitre, des phones du français et de l'anglais ont servi à réaliser le premier alignement du corpus d'apprentissage audio. Ces phones sont choisis de manière subjective, et sont sensés être proches des phones de l'amharique. L'alignement des données audio sert à ré-estimer ces modèles de départ. Les données audio sont alignées de nouveau et les modèles acoustiques composant le jeu de 33 phones sont ré-estimés. Ce processus est répété plusieurs fois jusqu'à ce que la vraisemblance des alignements, qui correspond

à la probabilité des observations lors des alignements n'augmente plus. Typiquement cinq ou six alignements sont nécessaires. Les modèles acoustiques obtenus sont normalement bien représentatifs des phones hors-contexte ou « monophones » de l'amharique.

Ensuite de nouveaux alignements sont réalisés pour modéliser les contextes des phones observés dans les transcriptions audio. Les modèles acoustiques obtenus sont des tri-phones à états liés avec 32 gaussiennes par état, 3 états par modèle, dépendants de la position inter-mot, c'est-à-dire que différents modèles sont utilisés pour des phones à l'intérieur des mots et pour des phones en frontière de mots. Les contextes peu observés sont regroupés à l'aide d'un arbre de décisions comme cela est décrit dans 1.4. Au total pour le système de référence, 10,7k contextes distincts sont modélisés par 8,5k états.

2.8 Modélisation linguistique

Les modèles de langage que nous avons utilisé sont issus de l'interpolation de deux modèles trigrammes ou quadrigrammes avec repli et lissage de type Kneser-Ney modifié [Kneser & Ney, 1995]. Le premier modèle est un modèle appris uniquement sur les transcriptions (240k mots) et le second sur les textes du Web (4,64M mots). La perplexité mesurée sur le corpus de développement (14,1k mots) est relativement élevée : $ppl = 372$ avec un taux d'OOV très élevé de 7,0%. Le tableau 2.13 donne le nombre de n -grammes pour $n = 1, 2, 3, 4$ du modèle quadrigramme utilisé dans le décodeur. À titre d'information, l'un modèle de langage utilisé au LIMSI pour la transcription d'émissions radio et télévision en anglais américain appris sur de très grands corpus de textes, comporte 65k unigrammes, 10M de bigrammes, 30M de trigrammes et 28M de quadrigrammes. D'une manière générale, pour une langue bien dotée, le nombre de bigrammes est de quelques millions d'unités, et les nombres de trigrammes et quadrigrammes sont de l'ordre de quelques dizaines de millions d'unités.

<i>n</i> -grammes	Nombre
1-g	133k
2-g	2163k
3-g	3615k
4-g	3859k

TAB. 2.13 – Nombre de n -grammes d'un modèle de langage quadrigramme utilisé pour la reconnaissance de l'amharique.

2.9 Évolution des performances en fonction de la quantité de données

Dans cette section, nous étudions l'évolution des taux d'erreurs de mots en fonction de la quantité de transcriptions utilisée pour estimer les modèles acoustiques et la partie transcriptions des modèles de langage.

La figure 2.5 donne l'évolution du nombre de contextes modélisés, en fonction de la taille en heures du corpus audio d'apprentissage pour des tailles de corpus entre dix minutes et dix heures de données. La figure montre également le nombre de mots des transcriptions correspondant aux différents corpus. Sans corpus d'apprentissage (point de départ), les modèles sont les 36 modèles « seed ». Avec dix minutes d'audio, qui correspondent à 1,5k mots de transcriptions, 140 contextes sont rencontrés. Avec dix heures de données soit 77k mots, ce nombre s'élève à 3,4k contextes. Avec la totalité des données (35h14min d'apprentissage), 10k contextes sont modélisés. L'évolution du nombre de modèles ressemble à une courbe logarithmique en fonction de la quantité de données.

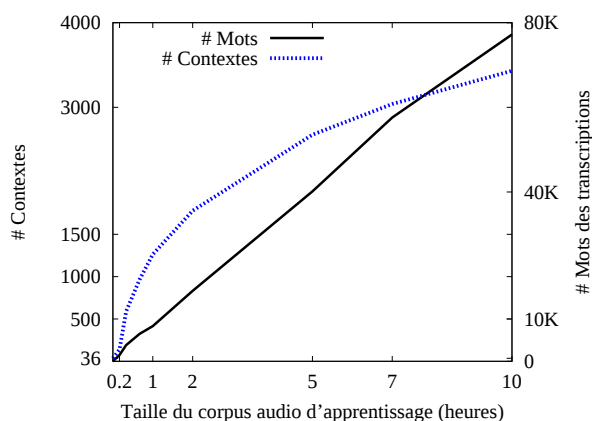


FIG. 2.5 – Nombre de contextes modélisés en fonction de différentes tailles de corpus d'apprentissage audio : 10, 20, 40, 60, 120, 300, 420, 600 minutes.

Pour chaque point de la courbe 2.5, un système de reconnaissance a été construit. Tous les systèmes ont la même architecture, la seule différence est la taille du corpus d'apprentissage utilisé pour entraîner les modèles acoustiques et estimer les modèles de langage. Les tailles des corpus utilisés sont : 10, 20, 40, 60, 120, 300, 420, 600 minutes ainsi que le corpus d'apprentissage dans son intégralité (35h). La figure 2.6 montre les performances de ces systèmes mesurés en taux d'erreurs de mots sur le corpus de développement (2h de parole, 14,1k mots). Pour chaque point de la courbe, le modèle de langage et le lexique utilisés ont été ré-estimés en utilisant uniquement les transcriptions manuelles qui ont servi à entraîner les modèles acoustiques.

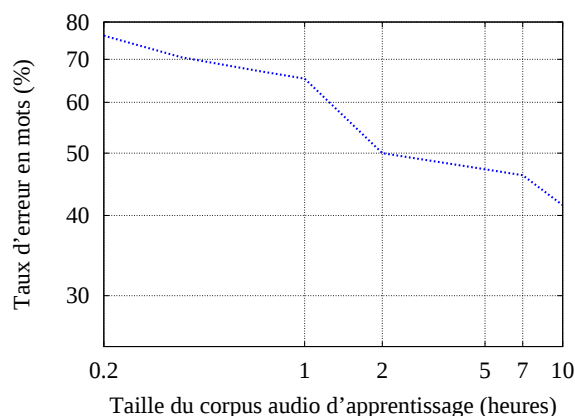


FIG. 2.6 – *Taux d'erreurs en mots en fonction de la taille du corpus de transcriptions manuelles.*

La courbe montre une diminution rapide du taux d'erreurs entre 10 minutes et 2h de données, le taux passe de plus de 76% à environ 50%. À partir de 2h de données, le taux d'erreurs diminue moins rapidement pour atteindre 41,5% avec 10h d'apprentissage. La pente forte entre 1h et 2h est due à l'amélioration des modèles acoustiques. En effet une expérience complémentaire a montré que le même taux d'erreurs a été obtenu avec un modèle de langage correspondant à 2h de données et des modèles acoustiques appris sur 1h de données. Enfin lorsque la totalité des corpus audio et textes d'apprentissage (respectivement 35h de parole transcrite totalisant 240k mots et 4,6M de mots de textes) sont utilisés, le taux d'erreurs est de 24,4%.

2.10 Vaut-il mieux collecter des textes ou transcrire quelques heures de données audio ?

Pour mieux identifier l'impact des modèles acoustiques et des modèles de langage sur le taux d'erreurs en fonction de la quantité de données, différents systèmes ont été construits. Quatre jeux de modèles acoustiques ont été construits avec 2 heures, 5 heures, 10 heures et 35 heures de données. Avec chacun des jeux de modèles acoustiques, différents modèles de langage (ML) sont utilisés. Ces ML diffèrent par la quantité de textes issus du Web utilisée pour estimer ces modèles : 10k, 100k, 500k, 1M et 4,6M mots (totalité des textes Web). Chaque ML est le résultat de l'interpolation d'un ML estimé sur l'un des cinq corpus de textes différents, avec un ML estimé sur les transcriptions correspondant aux sous-corpus d'apprentissage des modèles acoustiques.

Sélection des transcriptions

Comme précisé dans la section 2.3, les émissions transcrites couvrent une période d'un an, entre janvier 2003 et janvier 2004. Le corpus de devtest contient les transcriptions des émissions les plus récentes, datant des mois de décembre 2003 et janvier 2004. Pour le plus petit corpus de transcriptions, deux heures parmi les 35h totales de données ont été choisies en fonction des dates des émissions correspondantes. Ces deux heures datent du mois de novembre 2003, précédant d'un mois celles du corpus de devtest. Pour les corpus de 5h et 10h de transcriptions, aux deux heures du plus petit corpus ont été ajoutées le nombre d'heures adéquat, soit respectivement trois et huit heures qui proviennent d'émissions antérieures en date. Enfin, le corpus d'apprentissage total contient 35 heures de données.

Les corpus de 2h, 5h, 10h et 35h de transcriptions totalisent respectivement 17k, 35k, 70k et 240k mots.

Sélection des corpus de textes

La totalité des textes collectés sur le Web représente 4,6 millions de mots. Un premier texte de 10k mots est extrait en prenant une phrase sur 470 de ce corpus. Pour le corpus de 100k mots, nous avons ajouté au premier texte une phrase sur 51 du corpus initial, pour totaliser 100k mots. Et ainsi de suite pour les corpus plus grands, le corpus plus petit est conservé et sont ajoutées des phrases en nombre voulu pour atteindre la quantité de mots choisie. Nous disposons uniquement d'un seul fichier regroupant tous les textes collectés, sans information qui aurait permis de sélectionner des sous-corpus de texte par date et par source.

Taille des lexiques

Le tableau 2.14 donne les tailles des lexiques en nombre de mots pour chaque configuration de test. Les lignes du tableau correspondent aux trois corpus de transcription de 2h, 5h, 10h et 35h. Les nombres de mots et les nombres de mots distincts, communément appelés tokens et types, sont précisés. Les vocabulaires ont été construits simplement en prenant les mots distincts des transcriptions et les mots distincts des textes Web. Aucune limite inférieure d'occurrence n'a été utilisée sauf pour le corpus de textes Web le plus grand (4,6M de mots), où uniquement les mots apparaissant au moins trois fois ont été retenus. Pour les textes issus des transcriptions, aucune limite n'a été utilisée quelle que soit la quantité de mots.

Avec les plus petits corpus, à savoir 2h de transcriptions et 10k mots de textes Web, on compte seulement 11k mots distincts. Le nombre de mots distincts augmente très rapidement avec la taille des textes et des transcriptions, ce qui est caractéristique des langues à morphologie riche [Kirchhoff & Sarikaya, 2007]. Par exemple, pour 100k mots

de textes Web et 35h de transcriptions qui contiennent 240k mots, le lexique comprend presque 69k mots distincts. Les tailles des lexiques construits avec la totalité des textes Web sont plus petites que celles des lexiques construits avec 1M de mots. Cela est dû à la limite inférieure de trois occurrences appliquée pour la totalité du corpus Web. Le lexique correspondant à 1M de mots de textes Web et 35h de transcriptions contient 163k mots environ contre 133k avec 4,6M de mots des textes Web. Un seuil de deux occurrences minimum, qui consiste à soustraire les mots n'apparaissant qu'une seule fois, diminue de plus de 50% la taille du lexique. Sans contrainte sur le nombre d'occurrences pour les textes Web, les textes et les transcriptions contiennent 350k mots distincts.

<i>Transcriptions</i>		<i>Mots (textes)</i>				
<i>Heures</i>	<i>Mots/Types</i>	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	<i>17k / 7k</i>	11k	36k	96k	142k	114k
<i>5h</i>	<i>35k / 12k</i>	16k	39k	98k	144k	116k
<i>10h</i>	<i>70k / 21k</i>	24k	45k	103k	148k	119k
<i>35h</i>	<i>240k / 50k</i>	52k	69k	120k	163k	133k

TAB. 2.14 – *Taille des lexiques (nombre de mots distincts) pour chaque configuration de test. Les tailles sont données en fonction des tailles de textes Web (données en nombre de mots) et transcriptions (données en nombre d'heures et en nombre de mots et de types).*

Taux de mots hors-vocabulaire (mots OOV)

Le tableau 2.15 donne les taux de mots (tokens) hors-vocabulaire (OOV) pour chaque configuration. Pour 2h de transcriptions, avec un corpus de 10k mots de textes Web, 30,6% des mots du corpus devtest sont hors-vocabulaire. Avec un corpus de 4,6M de mots, ce taux tombe à 8,1%, et avec 35h de transcriptions, ce taux atteint sa valeur minimale de 7,0%.

<i>Transcriptions</i>		<i>Mots (textes)</i>					
<i>Heures</i>	<i>Mots/Types</i>	<i>0</i>	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	<i>17k / 7k</i>	36,2	30,6	18,8	11,5	9,0	8,1
<i>5h</i>	<i>35k / 12k</i>	28,5	25,7	17,3	11,1	8,8	7,9
<i>10h</i>	<i>70k / 21k</i>	22,7	21,4	15,7	10,6	8,4	7,7
<i>35h</i>	<i>240k / 50k</i>	14,5	13,9	12,1	9,1	7,5	7,0

TAB. 2.15 – *Taux de mots OOV (%) en fonction des tailles de textes Web (données en nombre de mots) et transcriptions (données en heures).*

La figure 2.7 illustre le tableau 2.15 en donnant l'évolution du taux de mots OOV en fonction de la taille des corpus de textes issus du Web. Trois courbes différentes sont

tracées pour les trois corpus de transcriptions de 2h (courbe noire), 10h (courbe verte avec point croix) et 35h (courbe rouge avec point étoile). La courbe du corpus 5h n'est pas tracée pour simplifier la figure, les taux d'OOV pour ce corpus sont dans le tableau. Les points de départ des courbes correspondent aux modèles de langage estimés uniquement sur les trois corpus de transcriptions.

À nombre de mots des textes Web constant, passer de 2h à 10h de transcriptions consiste à passer d'un corpus de 17k mots à un corpus de 70k mots de transcription. La diminution du taux d'OOV est de 11,8% en absolu lorsque l'on passe de 10k à 100k mots avec 2h de transcriptions soit 1,3% par ajout de 10k mots en moyenne. La diminution du taux d'OOV est de 9,2% absolus lorsque l'on passe de 2h à 10h de transcriptions soit 1,7% par ajout de 10k mots en moyenne. L'ajout de 10k mots de transcriptions est donc légèrement plus efficace que 10k mots de textes pour les très petits corpus, ce qui peut s'expliquer par la plus grande proximité en date et en style par rapport aux données de test. Cette différence tend cependant à diminuer avec la quantité de textes Web.

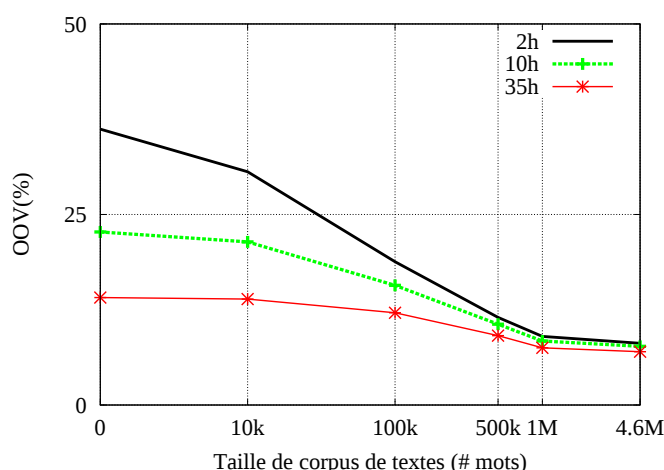


FIG. 2.7 – Taux de mots OOV en fonction de la taille en mots des textes collectés sur le Web. Les trois courbes correspondent aux trois corpus de transcription utilisées de 2h, 10h et 35h.

Perplexités

Les perplexités données dans cette section ont toutes été mesurées sur le corpus de dev-test de 14k mots. Nous avons vu dans la section 1.2 que des mesures de perplexité sont comparables uniquement si les lexiques associés aux ML comparés sont de taille identique. Pour pouvoir comparer les ML des différentes configurations, il est nécessaire de normaliser toutes les mesures avec une taille de lexique idéal estimée. Nous donnons néanmoins deux tableaux pour illustrer les valeurs de perplexités, avec et sans normalisation.

<i>Transcriptions</i>		<i>Mots (textes)</i>				
<i>Heures</i>	<i>Mots/Types</i>	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	<i>17k / 7k</i>	1279	1526	1317	1218	437
<i>5h</i>	<i>35k / 12k</i>	1337	1429	1198	1102	402
<i>10h</i>	<i>70k / 21k</i>	1344	1347	1122	1033	383
<i>35h</i>	<i>240k / 50k</i>	1181	1162	993	912	372

TAB. 2.16 – *Perplexités mesurées sur le texte de devtest pour chaque configuration de test. Aucune normalisation n'a été appliquée.*

Le tableau 2.16 donne les perplexités non-normalisées obtenues avec les différents ML. La perplexité mesurée avec le ML le plus grand, issu de l'interpolation du corpus entier de 35h de transcriptions et de 4,6M de mots de textes issus du Web, s'élève à 372. Notons que globalement, les perplexités diminuent avec l'augmentation de la quantité de textes ou de transcriptions utilisées pour estimer le ML, ce qui est une tendance qui semble naturelle. Il y a cependant une « anomalie » pour la configuration *2h/10k* pour laquelle la perplexité qui atteint 1279 est plus petite que les perplexités des configurations comprenant plus de mots comme *2h/100k*, *2h/500k*, *10h/10k* et *10h/100k*.

Il semble étonnant à première vue d'obtenir une perplexité avec le ML estimé sur 10h plus grande que celle obtenue avec le ML estimé sur 2h de transcriptions. En réalité ce résultat peut être dû au très fort taux d'OOV qui biaise la mesure de perplexité pour le modèle correspondant à 2h de transcriptions. En voici une explication : un mot hors-vocabulaire est associé à la balise `<UNK>` pour « unknown ». Pour toutes les configurations utilisées ici, tous les mots distincts (types) sont dans les différents lexiques sauf pour les trois configurations qui comprennent la totalité des textes Web (4,6M de mots), pour lesquelles les mots apparaissant au moins trois fois dans les textes ont été sélectionnés. Mis à part ces trois configurations, les autres fonctionnent à vocabulaire dit fermé, où il n'y a théoriquement aucun mot hors-vocabulaire dans les textes et transcriptions d'apprentissage des modèles de langage. Pour la création des ML à vocabulaire fermé, la boîte à outils de SRI [Stolcke, 2002] ajoute automatiquement le mot `<UNK>` au lexique et lui attribue la probabilité d'un singleton (un singleton est un mot qui n'apparaît qu'une seule fois dans les corpus de textes). Si les textes d'apprentissage comportent des mots hors-vocabulaire alors tous ces mots sont associés au mot unique `<UNK>` et la probabilité associée est calculée selon sa fréquence comme s'il s'agissait d'un mot normal.

Malgré l'étape de normalisation des textes, des mots mal normalisés, accolés à des chiffres ou des nombres par exemple, sont présents dans les textes d'apprentissage. Ces mots n'ont pas été sélectionnés pour construire les lexiques, ils sont donc hors-vocabulaire. Leur nombre étant supérieur à 1, la probabilité associée au mot `<UNK>` est donc plus grande que celle d'un singleton. Le tableau 2.17 donne à titre d'illustration les log-probabilités du mot `<UNK>` et des singletons pour les configurations respectives *2h/10k* et *10h/10k* estimées dans les ML correspondants. Plus la log-probabilité est petite, plus la perplexité

est grande. Le nombre de mots hors-vocabulaire est bien plus élevé pour le plus petit ML $2h/10k$ ce qui diminue artificiellement la perplexité mesurée avec ce modèle.

De plus, avec ces modèles de langage très petits, le repli aux unigrammes pour calculer la probabilité d'un n-gramme est fréquent. Les probabilités des singletons sont plus petites pour le corpus $10h/10k$ que pour le corpus $2h/10k$ ce qui contribuerait également à une perplexité plus grande obtenue avec le corpus d'apprentissage $10h/10k$.

On peut remarquer que dans les tous cas, la probabilité de la balise $\langle \text{UNK} \rangle$ est plus grande que celle des singletons. La perplexité d'un segment de phrase peut paraître très correcte, alors qu'en réalité le modèle de langage prédit beaucoup de mots inconnus, assimilés à la balise $\langle \text{UNK} \rangle$.

Configuration	10k/2h	10k/10h
$\langle \text{UNK} \rangle$	-3,9	-4,6
Singletons	-4,4	-5,0

TAB. 2.17 – Log-probabilités du mot $\langle \text{UNK} \rangle$ et des singletons dans les ML des configurations $10k/2h$ et $10k/10h$.

Nous avons vu que sans contrainte de fréquence minimale, le nombre de mots distincts s'élève à environ 350k mots en tout. Nous avons choisi une taille de 500k mots pour normaliser les perplexités, dont les valeurs sont reportées dans le tableau 2.18. Avec le ML le plus petit ($2h/10k$ mots), la perplexité normalisée est très élevée, elle est de 52586. Avec le ML le plus grand ($35h/4,6M$ mots), la perplexité est réduite à 852. Il est intéressant de remarquer que même avec 4,6M de textes, passer de 2h à 10h de transcriptions, soit un ajout de 53k mots de transcriptions, diminue la perplexité de 17% relatifs seulement, de 1149 à 955. L'ajout de transcriptions est de manière assez naturelle plus efficace que l'ajout de textes, du fait principalement de la plus grande proximité des thèmes abordés dans les transcriptions d'apprentissage et les transcriptions de test relativement proches en date, mais également du fait que les transcriptions ont une syntaxe de type « oral », différente a priori de celle des textes du Web.

Transcriptions		Mots (textes)				
Heures	Mots/Types	10k	100k	500k	1M	4,6M
2h	17k / 7k	52585	14726	5203	3545	1149
5h	35k / 12k	29902	11482	4500	3122	1033
10h	70k / 21k	17912	8977	3969	2801	955
35h	240k / 50k	6284	5000	2920	2218	852

TAB. 2.18 – Perplexités mesurées sur le texte de devtest pour chaque configuration de test. Une normalisation a été appliquée à ces valeurs en appliquant une taille de lexique virtuel de 500k mots.

Coefficients d'interpolation des ML

Le tableau 2.19 donne les coefficients d'interpolation des ML en fonction de la quantité de textes des transcriptions du corpus audio (2h, 5h, 10h et 35h) et de la quantité de textes provenant du Web (10k, 100k, 500k, 1M et 4,6M mots). Les coefficients correspondent au poids du ML des transcriptions. Ils sont optimisés sur le corpus de devtest (14,1k mots de transcriptions). Optimiser l'interpolation sur le corpus de devtest sur lequel sont testés et évalués les systèmes de reconnaissance présentés dans cette thèse revient à se placer dans des conditions particulièrement favorables.

Pour le ML estimé sur 2h de transcriptions, avec 10k mots de textes, le coefficient est très élevé, il vaut 0,71. Les 2 heures de transcriptions modélisent naturellement mieux les textes des transcriptions devtest que les 10k mots des textes Web. Le coefficient diminue très vite avec l'augmentation de la quantité de textes Web pour atteindre 0,2 environ à partir d'un million de mots. Plus la taille du corpus de textes est grande, plus le coefficient du ML estimé sur les transcriptions est petit. Ce comportement est vérifié pour les ML estimés sur 2h, 10h et la totalité du corpus de transcriptions. Enfin à taille de corpus de textes Web égale, plus la quantité de transcriptions est grande, plus le coefficient est grand. Avec la totalité des textes soit 4,6 millions de mots, les coefficients sont de 0,22, 0,28, 0,33 et 0,43 pour les modèles entraînés avec 2h, 5h, 10h et 35h respectivement.

<i>Transcriptions</i>	<i># Mots (textes)</i>				
	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	0,71	0,40	0,27	0,23	0,22
<i>5h</i>	0,84	0,53	0,38	0,33	0,28
<i>10h</i>	0,90	0,63	0,46	0,40	0,33
<i>35h</i>	0,95	0,81	0,61	0,52	0,43

TAB. 2.19 – Coefficients du ML estimé avec 2h, 5h, 10h et 35h de transcriptions pour l'interpolation des ML transcriptions/textes.

Comparaison des performances

Pour chaque configuration un système de reconnaissance a été construit et évalué sur le corpus de devtest. Tous les systèmes ont la même architecture en deux passes successives avec une adaptation non-supervisée des modèles acoustiques après la première passe [Gauvain *et al.*, 2002]. Des modèles acoustiques spécifiques à chacun des corpus audio de 2h, 5h, 10h et 35h ont été construits. Ces modèles sont tous des modèles de Markov (HMM) à états liés couvrant des contextes intra- et inter-mots, avec trois états par modèle et 32 gaussiennes par état. Les nombres de contextes modélisés et d'états des HMM sont donnés dans le tableau 2.20.

Le tableau 2.21 et la figure 2.8 illustrent les performances des différents systèmes en

<i>Transcriptions</i>	<i>2h</i>	<i>5h</i>	<i>10h</i>	<i>35h</i>
# <i>Contextes</i>	3027	4557	6323	10726
# <i>États liés</i>	1187	2286	3861	8554

TAB. 2.20 – Nombres de contextes modélisés et d'états des modèles acoustiques pour les sous-corpus audio (2h, 5h, 10h et 35h).

donnant pour les quatre systèmes construits avec 2h (courbe noire), 5h (courbe bleue), 10h (courbe verte) et 35h (courbe rouge) de données audio, les taux d'erreurs en mots en fonction de la taille en nombre de mots des textes Web utilisés pour estimer les ML.

La figure 2.8 montre une grande différence de performance entre le système 2h et le système 35h avec un taux d'erreurs environ deux fois plus grand pour le système 2h. Le système construit avec 35h de données présente un taux d'erreurs sur les mots (WER) variant de 31,4% à 24,4%, soit une différence de 22% relative avec respectivement 10k et 4,6M de mots de textes. Le système 2h présente un taux d'erreurs qui diminue de 25% relatifs entre 10k et 4,6M de mots de textes servant à estimer les ML. Il présente un taux d'erreurs de 48,5% soit une erreur tout les deux mots en moyenne.

Les performances du système 5h semblent peu différentes de celles du système 10h, avec des différences relatives valant 11%, 8%, 7%, 7% et 5% pour les quantités de textes dans l'ordre croissant. La courbe verte qui représente le WER du système 10h montre que les performances de ce système sont très proches du meilleur système, construit avec 35h de données audio. Pour 10k et 100k mots de textes, les différences de performances sont élevées entre les deux systèmes (22% et 16% relatifs respectivement). En revanche avec les corpus de textes plus grands (1M et 4,6M de mots), les performances sont proches, avec seulement 3% et 2% de différence absolue (10% et 8% relatifs). Ces chiffres montrent l'importance particulière des transcriptions audio pour la modélisation acoustique et linguistique lorsque les corpus de textes sont très petits, c'est-à-dire lorsque les quantités de textes et de transcriptions sont comparables.

Ces résultats semblent indiquer qu'à partir de textes d'un million de mots, les performances des systèmes 5h et 10h *a fortiori* sont relativement proches de celles du système 35h. Les 25h de transcriptions de différence entre ces deux systèmes apportent un faible gain. Cette conclusion suggère qu'à partir de 10h de données audio, collecter des textes serait plus efficace que transcrire des données audio supplémentaires.

Enfin un système construit avec 2h de données audio atteint des performances bien plus faibles que les systèmes 5h, 10h et 35h.

<i>Transcriptions</i>	<i># Mots (textes)</i>				
	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	64,4	59,6	55,5	53,0	48,5
<i>5h</i>	45,5	38,7	33,1	30,9	28,0
<i>10h</i>	40,4	35,5	30,9	28,7	26,7
<i>35h</i>	31,4	29,9	27,0	25,7	24,4

TAB. 2.21 – Taux d’erreurs en mots (WER en %) pour toutes les configurations audio : 2h, 5h, 10h et 35h, et textes : 10k, 100k, 500k, 1M, 4,6M de mots.

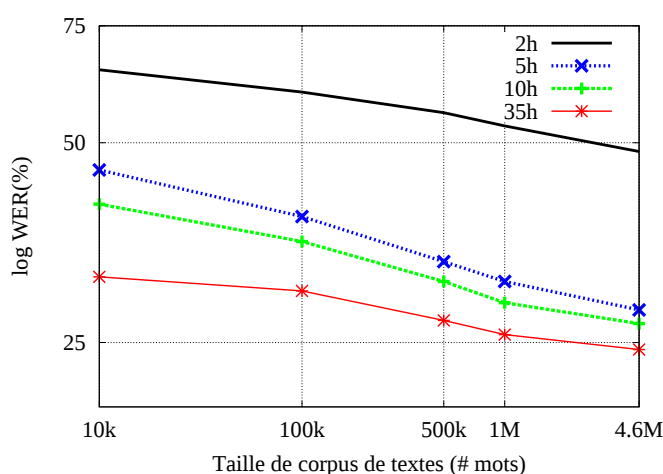


FIG. 2.8 – Taux d’erreurs en mots (WER en %) en fonction de la taille en mots des textes collectés sur le Web. Quatre courbes sont tracées pour les quatre corpus de transcriptions : 2h (courbe noire), 5h (courbe bleue), 10h (courbe verte) et 35h (courbe rouge).

Influences respectives du ML et des modèles acoustiques (MA)

La figure 2.8 a montré de nettes différences de performances entre le système 2h et le système 10h. D’autre part, nous avons vu plus haut que passer de 2h à 10h de transcriptions (ce qui correspond à un ajout de 53k mots) diminue fortement la perplexité normalisée des ML (17% relatifs).

Ces deux systèmes diffèrent de par les données audio (2h vs 10h) utilisées pour entraîner les modèles acoustiques et les transcriptions (17k vs 70k mots) utilisées pour estimer les modèles de langage. Pour évaluer si l’une de ces différences a un impact plus important que l’autre sur les performances, un système « croisé » a été construit. Les modèles acoustiques de ce système noté « MA2h/ML10h » ont été entraînés sur les 2h de transcriptions du système 2h, en revanche le modèle de langage a été estimé sur les 10h de transcrip-

tions du système 10h. De la même manière, un système « MA2h/ML35h » et un système « MA10h/ML35h » ont été construits.

Le tableau 2.22 et les figures 2.9 et 2.10 rappellent les performances des systèmes 2h, 10h et 35h et donnent les performances des systèmes croisés.

Quelle que soit la quantité de textes utilisée, les performances du système 10h sont bien meilleures que celles des systèmes 2h, MA2h/ML10h et MA2h/ML35h. Les gains observés en augmentant la taille du corpus de transcriptions sont plus élevés lorsque les textes sont en petite quantité. Avec un texte de 10k mots, le système MA2h/ML10h présente un gain relatif de 10% et le système MA2h/ML35h un gain relatif de 15% par rapport au système 2h. Avec le corpus entier de textes de 4,6M de mots, les gains relatifs sont identiques pour les deux systèmes croisés MA2h/ML10h et MA2h/ML35h, et sont de 4%. Ceci constitue un exemple de deux systèmes qui ont des perplexités normalisées différentes, 955 et 852 pour les systèmes MA2h/ML10h et MA2h/ML35h respectivement, mais qui présentent des performances de reconnaissance identiques.

Le système croisé AM10h/LM35h présente 16% de gains relatifs par rapport au système 10h avec 10k mots de texte. Avec des corpus de textes plus grands, les gains diminuent : ils sont respectivement de 8%, 4%, 2% et 0% avec 100k, 500k, 1M et 4,6M de mots. En revanche la différence de performance entre les systèmes AM10h/LM35h et 35h est constante en fonction de la quantité de textes. Le gain relatif du système 35h est constant : égal à 8%, il est donc dû vraisemblablement à de meilleurs modèles acoustiques.

<i>Transcriptions</i>	<i># Mots (textes)</i>				
	<i>10k</i>	<i>100k</i>	<i>500k</i>	<i>1M</i>	<i>4,6M</i>
<i>2h</i>	64,4	59,6	55,5	53,0	48,5
<i>MA2h/ML10h</i>	57,9	56,5	53,7	51,3	46,6
<i>MA2h/ML35h</i>	54,7	53,9	52,0	50,4	46,6
<i>10h</i>	40,4	35,5	30,9	28,7	26,7
<i>MA10h/ML35h</i>	34,0	32,6	29,6	28,1	26,6
<i>35h</i>	31,4	29,9	27,0	25,7	24,4

TAB. 2.22 – *Taux d'erreurs en mots (WER en %) pour toutes les configurations : transcriptions de 2h, 10h et 35h de données audio, textes de 10k, 100k, 500k, 1M, 4,6M de mots.*

De façon synthétique, les courbes MA2h/ML10h et MA2h/ML35h sont très proches de la courbe 2h, ce qui indiquerait qu'avec très peu de données audio (2h en l'occurrence), transcrire des heures supplémentaires est crucial pour améliorer les performances. Ajouter seulement des transcriptions pour estimer les ML améliore le WER de seulement 5% relatifs par rapport au système 2h, alors que l'ajout de ces heures pour le ML et les MA réduit le WER entre 30% et 40%, toujours en comparaison avec le modèle 2h.

Enfin le système MA10h/ML35h se comporte plus comme le système 35h avec de petites

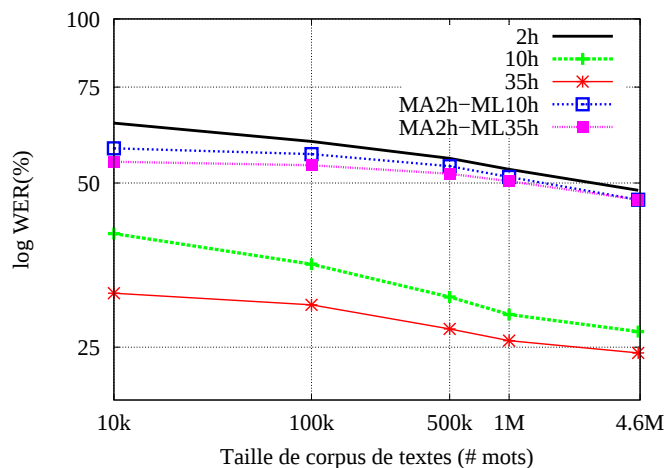


FIG. 2.9 – Log des taux d’erreurs en mots (WER en %) en fonction de la taille en mots des textes collectés sur le Web. La figure permet de comparer les systèmes 2h (noir), 10h (vert) et 35h (rouge) avec les systèmes croisés MA2h/ML10h (bleu) et MA2h/ML35h (violet).

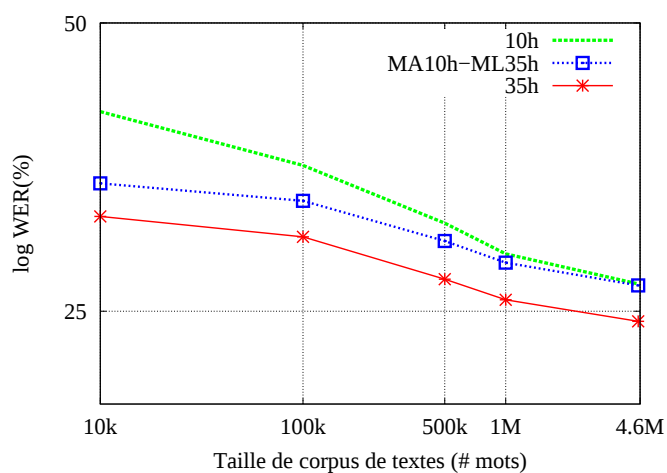


FIG. 2.10 – Zoom entre 20% et 50% en ordonnée de la figure 2.9. La figure reprend les courbes des systèmes 10h et 35h et montre celle du système croisé MA10h/ML35h (bleu).

quantités de textes (10k et 100k mots), et plus comme le système 10h avec plus de textes. Cela montre que l'influence de la quantité de transcriptions dans le ML est réduite lorsque des quantités plus importantes de textes sont utilisées. Remarquons cependant que le coefficient d'interpolation pour la composante des transcriptions est élevé, autour de 0,40.

2.11 Premières expériences de décomposition des mots

La taille du lexique et le fort taux de mots hors-vocabulaire peuvent être diminués en séparant des affixes, que ce soient de « vrais » affixes, c'est-à-dire des affixes pertinents si l'on étudie la morphologie de la langue, ou que ce soient des affixes « artificiels », trouvés par la machine et sans valeur linguistique. Obtenir automatiquement des affixes avec un sens linguistique est plus satisfaisant, néanmoins dans une ultime étape, les affixes sont ré-agglutinés aux radicaux en sortie de décodeur pour reconstituer des mots entiers, et les affixes n'apparaissent plus.

Dans cette partie, nous allons décrire les tout premiers résultats de reconnaissance obtenus avec des décompositions de mots. Les gains observés au cours de ces expériences nous ont encouragé à développer des méthodes de décomposition des mots.

Le choix des affixes

Détecter les affixes automatiquement présente l'avantage de ne pas utiliser de connaissances linguistiques spécifiques à la langue cible et rend la méthode portable à d'autres langues facilement. L'algorithme de Harris [Harris, 1955] est un algorithme de détection des frontières de morphèmes indépendant de la langue, nécessitant simplement un corpus de mots de la langue cible. Il sera décrit de manière approfondie dans la section 4.2.1. Il exploite le fait qu'un début de mot de k caractères a naturellement peu de caractères successeurs distincts possibles pour former des mots qui existent dans la langue traitée pour k suffisamment grand. Au rang $k+1$ ce nombre réduira davantage. Si ce nombre augmente subitement pour un début de mot de k caractères alors ce début de mot est un morphème candidat, pouvant se composer avec d'autres morphèmes commençant par des lettres distinctes variées. Ainsi l'algorithme compte le nombre de caractères successeurs distincts possibles pour tous les débuts de mots de taille k et propose des frontières de morphèmes pour ces mots lorsqu'un maximum local est trouvé. Il ne s'agit pas de réaliser une analyse morphologique mais de dégager quelques affixes potentiels les plus fréquents. Séparer ces affixes des mots du lexique permet de réduire le nombre de lexèmes et d'augmenter la représentation de certains n -grammes peu observés [Adda-Decker, 2003].

Une liste de sept affixes (cinq préfixes et deux suffixes) a été retenue pour les premières expériences rapportées ici. Les sept affixes sont les plus fréquents parmi ceux détectés par l'algorithme. Les affixes détectés par l'algorithme de Harris qui sont moins fréquents

que les sept retenus sont beaucoup moins fréquents. Les sept affixes sont séparés des mots dont la taille après séparation est d'au moins deux syllabes. Un signe « + » est accolé aux affixes pour pouvoir recombinaison les mots par la suite. Le tableau suivant donne la liste des affixes :

<i>préfixes (5)</i>		<i>suffixes (2)</i>	
አንደ+	?xnxdE+	+ኸው	+CEwx
አንዳ+	?xnxda+	+ኸ	+mx
አንዲ+	?xnxdI+		
አንድ+	?xnxdx+		
የ+	jE+		

TAB. 2.23 – *Affixes retenus (5 préfixes et les 2 suffixes). Chaque affixe est donné en script amharique et dans sa forme transcrite avec le jeu de caractères choisis par nous. Le point d'interrogation représente un coup de glotte.*

Le nombre de mots du lexique est réduit de plus de 11%, de 133k à 119k mots. Le taux de mots hors-vocabulaire (OOV) diminue de 7,0% à 4,8% soit une réduction absolue de 2,2%. Le taux de mots OOV reste néanmoins très élevé.

Résultats des expériences de reconnaissance

De nouveaux modèles acoustiques ont été appris pour la représentation avec affixes séparés puisqu'ils sont dépendants de la position intra et inter-mots. Un nouveau modèle de langage quadrigramme a été généré, estimé sur les textes et les transcriptions avec mots décomposés. Le système de reconnaissance est le même que celui qui a servi pour la représentation en mots entiers, seuls les modèles acoustiques, le modèle de langage et le lexique de prononciation différent. Les résultats suivants ont été obtenus avec 10,6k modèles acoustiques (8,7k états).

Le tableau 2.24 donne les taux d'erreurs pour la représentation avec affixes séparés et après recombinaison des mots. Le taux obtenu avec affixes séparés est nettement inférieur car le nombre de mots est plus grand avec cette représentation (17,3k mots) et les affixes sont globalement très bien reconnus. En recombinaison les affixes (grâce au signe « + » accolé), le taux d'erreurs augmente. Un gain relatif de 3,1% (gain absolu de 0,8%) est néanmoins observé par rapport au taux d'erreurs de 25,9% avec le système appris sur les mots entiers (appelé S_{mots} par la suite). Ce taux d'erreurs de référence est plus élevé que le taux de 24,4% obtenu dans les mêmes conditions et reporté ci-dessus, les paramètres du décodeur (poids relatifs des scores acoustiques/linguistiques) ayant été optimisés entre temps. Les paramètres sont identiques pour les systèmes de référence mots entiers et affixes séparés.

Le tableau 2.25 donne un exemple de sortie du décodeur où le système avec affixes

<i>Représentation</i>	<i>Taux d'erreurs de mots</i>
<i>Affixes séparés</i>	21,6%
<i>Mots recomposés</i>	25,1%
<i>Référence mots entiers</i>	25,9%

TAB. 2.24 – *Taux d'erreurs sur les mots avant et après recombinaison des affixes.*

<i>Système</i>	<i>Phrase</i>			<i>Log-v</i>
<i>S_{mots}</i>	?iraKxlajx		jESxgxgxrx	-9,5551
	?iraKx	lajx	jESxgxgxrx	-9,6559
<i>S_{affixes}</i>	?iraKx	lajx	jE+ Sxgxgxrx	-9,2613
	?iraKxlajx		jE+ Sxgxgxrx	-10,1367

TAB. 2.25 – *Exemple de phrase correctement reconnue par S_{affixes} mais erronée pour S_{mots}, comparaison des log-vraisemblances données dans la colonne Log-v. Les lignes S_{mots} et S_{affixes} donnent les sorties respectives des systèmes.*

séparés a été meilleur. Les phrases de référence, en gras dans le tableau, ont respectivement trois mots pour la représentation en mots entiers (*S_{mots}*) et quatre mots pour la représentation avec affixes séparés (*S_{affixes}*). Le système *S_{mots}* n'a pas bien reconnu la phrase, la sortie obtenue ayant deux mots au lieu de trois. En revanche le système *S_{affixes}* a correctement reconnu la phrase de référence. Le tableau donne, pour chaque système, la log-vraisemblance (log-v) pour la phrase correcte (phrase de référence) et pour la phrase erronée résultat du décodage par le système *S_{mots}*. Pour *S_{affixes}*, la phrase erronée est la phrase erronée de *S_{mots}* après séparation de l'affixe.

La vraisemblance, qui est la probabilité des suites de mots, est utilisée par le décodeur pour sélectionner la meilleure hypothèse. Les mots « lajx » et « jE+ » sont parmi les mots les plus fréquents des textes avec affixes séparés, alors que le mot « ?iraKxlajx » est beaucoup moins fréquent, ce qui favorise la phrase correcte, obtenue par *S_{affixes}*, qui contient « lajx fg » et « jE+ ». Pour le système *S_{mots}*, la forte probabilité de l'unigramme « lajx » ne suffit pas à favoriser la phrase de référence. Le mot « jESxgxgxrx » étant rare, la séparation de l'affixe « jE+ » a été bénéfique.

2.12 Conclusion

Dans ce chapitre, nous avons présenté la langue amharique, qui bien que faisant l'objet de recherches et de développement importants en reconnaissance de la parole, est une langue qui possède peu de textes disponibles sur Internet.

L'un des premiers problèmes que nous avons rencontré fut la création d'un lexique de

prononciations. Nous avons proposé une méthodologie de création du lexique de prononciations qui permet d'identifier des variantes de prononciation sans connaissance linguistique. À partir d'alignements phonotactiques autorisant la substitution de chaque phone par un autre, les substitutions les plus fréquentes sont sélectionnées. Le taux appelé « variant2+ », défini comme le pourcentage de mots qui n'ont pas été alignés avec la prononciation principale, permet de mesurer l'intérêt des variantes de prononciations proposées. Cette approche complète l'approche purement graphémique, qui est la méthode la plus simple lorsque l'on dispose pas de connaissance linguistique sur la langue, en proposant des variantes de prononciations qui peuvent améliorer la qualité des alignements, et donc améliorer les modèles acoustiques.

Dans un deuxième temps, l'influence des quantités de données audio transcrites et des quantités de texte a été étudiée de manière approfondie. Les performances de systèmes de reconnaissance standards entraînés sur des quantités de données audio semblent montrer qu'avec moins de dix heures de données audio transcrites, l'influence de la quantité des textes utilisés pour estimer les modèles de langage est moindre comparée à l'influence des modèles acoustiques. En revanche, cette tendance semble s'inverser à partir de dix heures de données d'apprentissage audio. Les différences de performances entre des systèmes entraînés sur 10h et sur 35h sont proches lorsqu'un corpus de textes d'au moins un million de mots est utilisé pour l'entraînement des modèles de langage. Ce résultat suggère qu'à partir de ce point de fonctionnement (10h de données audio transcrites et 1M de mots), collecter de nouvelles quantités de textes serait plus efficace que transcrire quelques heures de données audio supplémentaires. Cette interprétation est à nuancer dans la mesure où ces résultats ont été obtenus sur une seule langue. Généraliser à tous types de langues dans ces conditions n'est pas possible.

Enfin une première expérience de décompositions de mots qui a été menée en début de thèse, a été décrite en fin de chapitre. La séparation d'un jeu de sept affixes parmi les plus fréquents (cinq préfixes et deux suffixes), choisis de manière empirique, a montré un gain modeste mais significatif de 3% relatifs par rapport au système fondé sur les mots entiers. Ce gain a été le point de départ des recherches qui ont suivi sur la modélisation lexicale.

Chapitre 3

Modélisation lexicale

Ce chapitre présente différentes représentations lexicales pouvant être utilisées pour la reconnaissance de la parole et dresse un état de l'art sur ce thème. À l'exception de quelques langues comme le mandarin où ce sont les caractères qui sont utilisés comme unités de reconnaissance, le mot est en général l'unité de référence en reconnaissance de la parole. Le mot sert à évaluer les performances d'un système. Cependant, même pour les langues qui séparent les mots par un caractère (un espace en général), la question d'identifier les mots n'est pas triviale, et fait intervenir des choix de normalisation souvent différents selon les caractéristiques morphologiques de la langue traitée. Pour des langues peu dotées pour lesquelles une majorité de mots sont peu représentés, le mot est-il vraiment adapté ? Y aurait-il des unités plus petites plus pertinentes compte tenu du manque de données d'observation ? Ce chapitre constitue une introduction à ces questions en présentant différents niveaux de représentation lexicale. Les chapitres qui suivront exploiteront ces réflexions en mettant en pratique une recherche automatique d'unités lexicales adaptées à la reconnaissance de la parole.

3.1 Diversité des systèmes d'écriture

Daniels et Bright (1996) définissent un système d'écriture comme étant « un système de marques plus ou moins permanentes utilisées pour représenter une parole de façon à ce qu'elle puisse être reproduite plus ou moins exactement sans nécessiter l'intervention de l'émetteur ».

Dans cette définition, il y a l'idée que l'écrit doit représenter un énoncé oral. Ce n'est pas toujours strictement le rôle de l'écrit comme par exemple pour la littérature où le style langagier peut être très éloigné du style oral. Une définition qui fait intervenir l'idée de véhiculer un message, du sens, serait peut-être plus générale, comme celle de Katrin Kirchhoff dans le chapitre 2 de « Multilingual Speech Processing » [Schultz & Kirchhoff,

2006] : « un système d'écriture est un jeu de symboles accompagné de conventions permettant de communiquer un message avec des mots précis ». Apparaissent les notions de mot et de conventions sujettes à de nombreuses questions dont certaines relatives à la reconnaissance de la parole seront soulevées dans la section 3.5.

Les systèmes d'écriture se distinguent par le niveau de représentation des unités de base de ces systèmes, appelées « graphèmes ». Peter T. Daniels décrit trois systèmes différents :

- « Logosyllabaire » : les caractères représentent à la fois des morphèmes et à la fois des syllabes, c'est l'exemple du système hanzi chinois.
- « Syllabaire » : les symboles représentent des syllabes. Les systèmes Hirigana et Katakana du japonais en sont un exemple. L'amharique à la différence de l'arabe ou de l'hébreu et qui est pourtant une langue sémitique possède son propre syllabaire. Le terme « abugida » est parfois utilisé pour désigner les langues qui ont des symboles représentant des consonnes avec une voyelle par défaut (amharique et langues indiennes).
- « Alphabet » : les caractères représentent des phonèmes. Les langues occidentales comme les langues européennes (anglais, espagnol, français, portugais), le russe ou des langues sémitiques comme l'arabe et l'hébreu utilisent des alphabets. Le terme « abjad » est parfois utilisé pour désigner les langues qui ne représentent que les consonnes, les voyelles étant indiquées par des diacritiques uniquement pour désambiguïser lorsque c'est nécessaire (c'est le cas de l'arabe, de l'hébreu).

Daniels et Bright mentionnent un quatrième système appelé « featural » en anglais pour lequel certains caractères représentent les traits distinctifs des phonèmes, c'est l'exemple du système coréen hankul.

Sampson classe les systèmes d'écriture en deux grands types qui englobent ceux décrits ci-dessus [Sampson, 1985]. Il s'agit des systèmes dits « logographiques » où les symboles représentent des morphèmes et les systèmes dits « phonographiques » où les symboles sont liés aux unités sonores de base des sons de la langue. Les deux systèmes de type syllabaire et alphabet sont des systèmes phonographiques. Les systèmes logosyllabaires sont des systèmes logographiques. Les différents systèmes d'écriture peuvent être rencontrés pour une même langue dans des proportions variables ce qui empêche parfois de pouvoir classer un système de manière unique.

La figure 3.1 donne des exemples des différents systèmes. La figure 3.2 donne une carte des systèmes d'écriture dans le monde (http://en.wikipedia.org/wiki/Writing_system). On voit notamment que l'alphabet latin est très répandu, il est utilisé sur tous les continents sauf le continent asiatique.

မြစ် ငွေတူသည့်က ရလဒ်မှာ ငါတို့ ဘုံ၌ ခွဲလွှဲသော လူတွေက ငွေ ငါတို့ ငါ့

يولد جميع الناس أحراراً متساوين في الكرامة والحقوق. وقد وهبوا
عقلاً وضميراً وعليهم أن يعامل بعضهم بعضاً بروح الإخاء.

ቃለ በረከት ዘሄኖክ ዘከመ ባረከ ጎሩያነ ወጻድቃነ እለ ሀለው ይኩኑ
በዕለተ ምገዳቤ ለአሰሰሎ ኩሉ አኩያን ወረሲዓን።

人人生而自由，在尊严和权利上一律平等。他们赋有理性
和良心，并应以兄弟关系的精神互相对待。

FIG. 3.1 – Exemples de différents systèmes d'écriture (source omniglot.com) : trois systèmes phonographiques de haut en bas le thaï, l'arabe standard et le script geez ou ge'ez qui sert pour l'amharique entre autres et un système logographique, le système hàanzi chinois. Les systèmes phonographiques sont les systèmes où un symbole représente un phonème. Les systèmes logographiques sont les systèmes où un symbole représente un morphème.

3.2 Translittération versus transcription

Lorsqu'un nouveau système de reconnaissance est élaboré pour une langue donnée, une étape souvent nécessaire consiste à choisir un jeu de caractères distincts du jeu originel de la langue pour avoir une représentation des mots orthographiques simple à utiliser avec une machine. Le nom général de cette étape de normalisation que l'on trouve dans la littérature est « romanisation », ou « translittération » ou encore « transcription ». Quelles sont les différences entre ces acceptions ?

Le syllabaire amharique, par exemple, qui comprend un peu plus de 240 symboles, a été transcodé vers un jeu de 26 consonnes et 7 voyelles qui sont des caractères latins. Cette étape est appelée « translittération » ou « transcription » en fonction de la possibilité de revenir aux caractères originels de manière non-ambiguë. Si à chaque caractère originel

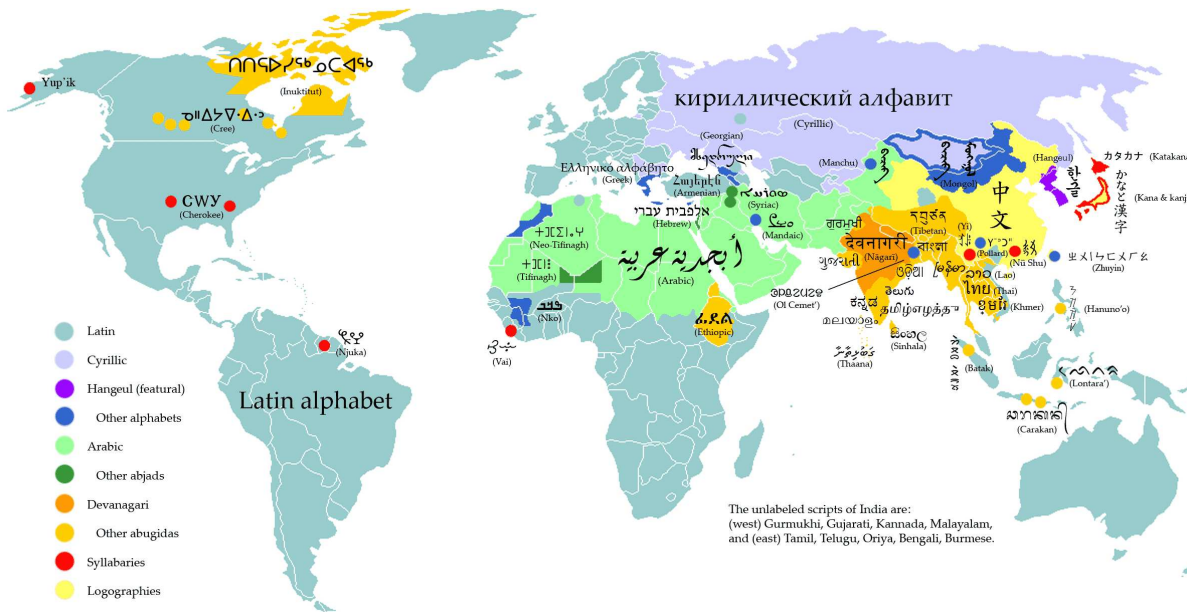


FIG. 3.2 – Répartition des différents systèmes d'écriture dans le monde (source : wikipedia.org).

correspond un unique caractère de conversion, la conversion est appelée translittération, sinon il s'agit d'une transcription.

Cette distinction a été définie par le comité technique international responsable de l'élaboration de standards pour la conversion des langues écrites. Ce comité appelé ISO/TC46/SC2 a produit des standards de transcription et de translittération avec des caractères latins pour des langues comme le grec, l'hébreu, l'arabe, le perse, le coréen, le thaï ou des alphabets comme l'alphabet cyrillique. John Clews, responsable de ce comité écrit dans [Clews, 1997] : « What is the difference between transliteration and transcription? Transliteration is the representation of letters of one script by the letters of another; transcription is the representation of sounds of one language in letters of one script ».

Toujours dans [Clews, 1997], Clews explique que la translittération est bien adaptée aux langues phonétiques puisque la sortie de la translittération peut être lue par quelqu'un qui ne connaît pas forcément les règles qui ont servi à aboutir au code translittéré :

« Transliteration can work very well for extremely phonetic languages and scripts like the Cyrillic, Greek, Armenian, and Georgian scripts in Europe, and most scripts of South and Southeast Asia, such as Devanagari, Panjabi, Gujarati, Bengali and Oriya; Kannada, Malayalam, Telugu, Tamil; Sinhala; Maldivian; Lao; Burmese and Khmer; and for Amharic (used in Ethiopia and Eritrea) ».

D'autres langues n'ont pas de lien simple entre écrit et oral et la transcription offre a priori une représentation plus lisible qu'une translittération, c'est le cas des idéogrammes chinois par exemple.

Pour des langues qui bénéficient d'un effort de recherche, il peut exister des conversions standards, utilisées par les communautés de traitement du langage. C'est le cas de l'amharique par exemple, avec le système SERA (System for Ethiopic Representation in ASCII) qui convertit les symboles du syllabaire avec un ou deux caractères latins pour chaque symbole. En amharique, il existe de nombreux caractères distincts homophones et les tables de conversion associent la même suite de lettres latines aux symboles homophones. Il n'est donc pas toujours possible de retrouver le symbole initial à partir de la forme latinisée ou romanisée puisqu'il y a le choix entre les différents symboles homophones. La conversion des caractères utilisés pour l'amharique est donc plutôt une transcription qu'une translittération.

3.3 Conversion graphème-phonème

Comme nous l'avons vu au chapitre 1, à chaque entrée lexicale doit correspondre une ou plusieurs prononciations qui sont associées à une séquence de modèles acoustiques, pour faire le lien entre les modèles acoustiques et les modèles de langage. L'entrée lexicale est par nature *graphémique*, et la prononciation est *phonémique*, si les modèles acoustiques représentent des phones. Le lien entre l'entrée lexicale et les prononciations associées est couramment appelé *conversion graphème-phonème* (GP). La conversion GP trouve de très nombreuses applications, dans les systèmes de reconnaissance de la parole et les systèmes de synthèse vocale, largement utilisés de nos jours.

La difficulté d'établir une conversion graphème-phonème varie considérablement en fonction de la langue étudiée. Ce lien peut être très simple et direct pour certaines langues (espagnol, turc, amharique...) [Killer *et al.*, 2003]. L'écrit amharique, par exemple, est une sorte de transcription simple de l'oral [Yacob, 2003], ainsi la génération des prononciations à partir des mots orthographiques ne pose pas de problème particulier. Pour d'autres langues comme le français par exemple, la conversion est moins directe. Il y a des suites de plusieurs lettres qui représentent un phonème unique, de nombreuses lettres ou suites de lettres homophones, par exemple *o*, *au* et *eau*. Un autre exemple de problème très courant en français sont les mots qui finissent en *-ent*. Cette terminaison est une désinence très courante pour la troisième personne du pluriel des verbes, quelque soit le temps de conjugaison. Elle est atone pour les verbes mais elle est prononcée /ã/ pour beaucoup d'autres mots, comme *lent*, *dent* par exemple. Il est toute fois possible de trouver les prononciations des mots à partir d'une liste de règles. Le système de conversion GP GRAPHON+, par exemple, utilise quelques deux mille règles, et donne un taux d'erreurs sur les mots inférieur à 1%, testé sur un dictionnaire de 270k entrées [Boula de Mareüil, 1997].

Enfin pour d'autres langues comme l'anglais par exemple, un graphème peut avoir de nombreuses prononciations très différentes en fonction du contexte et le nombre d'exceptions est très grand [Schultz & Kirchhoff, 2006]. Dans ce cas, l'élaboration d'un lexique de prononciations est principalement manuelle. Il faut également signaler les problèmes spécifiques qui peuvent se poser pour des langues qui omettent à l'écrit une partie de l'information présente à l'oral. L'absence des voyelles (arabe, hébreu), ou des tons à l'écrit, sont des exemples. Il peut être bénéfique de retrouver cette information perdue, c'est le cas en reconnaissance de la parole en arabe standard, où l'utilisation des formes voyellées des mots a mené à de légers gains de performance [Messaoudi *et al.*, 2006].

Pour conclure, il est important de faire la distinction entre modélisation lexicale, qui concerne les unités graphémiques de la langue, et modélisation acoustico-phonétique, qui concerne les modèles acoustiques utilisés pour la reconnaissance. Cependant ces deux types de questions sont très liées. Les travaux menés au cours de cette thèse sont plus orientés sur la modélisation lexicale, et nous nous sommes en quelque sorte affranchis des difficultés posées par la génération des prononciations en choisissant des langues qui ont une conversion graphème-phonème simple.

3.4 Quelques éléments de morphologie

La morphologie est une branche de la linguistique qui étudie la formation et la structure des mots. Le but modeste de ce paragraphe est de tenter de cerner les différences générales entre ce qu'on appelle des mots, des lexèmes et des morphèmes, pour ensuite fuir les difficultés en utilisant le terme « morphe », qui désigne toute sous-unité de mot et éventuellement le mot lui-même.

Pour l'instant, considérons que les mots sont des unités séparées par un espace (pour les langues qui ont un séparateur de mots). Des problèmes relatifs à la délimitation des mots seront plus amplement discutés dans le paragraphe 3.5. Un premier concept important est celui de lexème, qui désigne un concept abstrait regroupant plusieurs mots (par exemple ALLER pour aller, vais, vas, allé, etc...), liés par un sens général commun. Le fait de mettre le lexème en majuscules est simplement une convention que l'on peut adopter ou pas, pour le différencier du mot. Les morphèmes, quant à eux, désignent les unités les plus petites qui constituent les mots, qui font sens. Ce sont des unités abstraites qui peuvent exprimer un trait grammatical (marque du pluriel ou du genre par exemple), ou un trait lexical. Par exemple, le verbe *aller* pourrait être vu comme le morphème *all-* et le morphème *-er*, marque de l'infinitif des verbes du premier groupe. Le tableau ci-dessous donne quelques exemples simples, pour illustrer ces définitions.

Mots	aller, allé, vais, vas. . .
Lexème	ALLER
Morphèmes	all -er, all -é, vai -s, va -s. . .

	Français	Russe	Arabe	Finnois
Nombre moyen de mots par lexème	5	21	6800	2000(N) 14000(V)

TAB. 3.1 – Nombre moyen de mots par lexème mesuré sur corpus (français, russe) et estimé (arabe, finnois). Ces chiffres sont issus de [Kirchhoff & Sarikaya, 2007]

Il existe de multiples façons de créer un nouveau mot, par exemple en intégrant un mot de langue étrangère ou en créant un néologisme. Deux processus principaux sont classiquement mis en avant en morphologie [Matthews, 1991]. Il s’agit de la dérivation et de la composition. La dérivation consiste à former un mot à partir d’éléments dont l’un au moins n’a pas d’existence autonome comme mot. La dérivation change en général la catégorie syntaxique du mot concerné. L’ajout d’un suffixe est un exemple de dérivation. En français par exemple, le suffixe *-ment* peut compléter un adjectif pour donner un adverbe, ainsi *visible* donne *visiblement*. En anglais, le mot *unguardedly* peut être décomposé en *unguarded+ly* (et même en *un+guard+ed+ly*), et le mot *overcritical* par *overcritic+al* (en *over+critic+al* également).

La composition associe deux mots complets, leur forme étant éventuellement altérée, pour former un mot nouveau. Le français qui n’est pas une langue compositionnelle ajoute souvent un tiret entre les deux mots, comme par exemple dans les mots *tire-bouchon* ou *pare-chocs*. Dans les dictionnaires, le tiret n’est pas considéré pour établir l’ordre lexicographique. En anglais, où l’utilisation du tiret de séparation de mots composés est très libre, on peut trouver par exemple les mots *firefighter* et *timeconsuming* mais aussi *fire-fighter* et *time-consuming*.

Les langues ont des morphologies plus ou moins complexes. Par exemple, les langues dites à morphologie riche sont les langues pour lesquelles il y a beaucoup de mots par lexème. Pour ces langues, le nombre de mots distincts augmente très fortement avec le nombre de mots d’un corpus de textes. Le tableau 3.1 donne les nombres moyens de mots par lexème mesurés sur corpus pour le français et le russe, et estimés pour l’arabe et le finnois [Kirchhoff & Sarikaya, 2007]. Les valeurs avancées dans cette étude sont très élevées pour l’arabe et le finnois, et correspondent sans doute à une estimation sur-évaluée. Néanmoins, elles reflètent le processus de formation des mots de ces deux langues basé sur la flexion et la composition d’un mot racine qui est dérivé pour préciser des sens nouveaux. Ce sont des langues à morphologie riche. Pour ces langues, la modélisation de la morphologie est très importante, et est un point très étudié en traitement automatique, en particulier pour la reconnaissance de la parole.

3.5 Le mot : unité lexicale de base en reconnaissance automatique

Le mot est généralement l'unité de référence en reconnaissance de la parole. Si l'on rappelle l'équation de base de la reconnaissance de la parole $\hat{M} = \operatorname{argmax}_M P(M|S)$, donnée et développée dans la section 1.1, le but est de trouver la meilleure séquence de mots à partir du signal S . Les systèmes de reconnaissance sont donc en général évalués avec une mesure d'erreur moyenne sur les mots, appelée WER pour Word Error Rate. Rappelons la définition donnée à la section 1.6 : il est défini par la somme de trois types d'erreurs qui sont l'insertion $\#I$, la substitution $\#S$ et l'élision $\#D$ de mots, moyennée par le nombre de mots N de la référence : $(\#I + \#S + \#D)/N$. Ce taux est calculé après alignement dynamique de l'hypothèse du décodeur avec une référence et d'après sa définition il peut être supérieur à 100%.

En linguistique, la notion générale de *mot* est souvent décrite comme problématique, des difficultés apparaissant lorsque l'on veut délimiter les mots. André Martinet, dans « Éléments de linguistique générale » (1980), préfère associer au mot la notion de syntagme autonome, dans lequel des éléments élémentaires (appelés *monèmes*...), qui sont définis comme « étroitement unis par le sens », ne sont pas séparés. Ainsi il considère par exemple, le syntagme *sac à main* comme un seul mot.

L'utilisation de tels syntagmes paraît intéressante pour des applications de traitement automatique du langage naturel comme par exemple la traduction automatique. En reconnaissance de la parole, ce sont les mots graphiques qui sont utilisés, c'est-à-dire que sont considérés comme des mots différents, des chaînes de caractères séparées par un caractère séparateur de mots, qui est en général l'espace. Néanmoins la définition des mots du lexique de reconnaissance ne fait pas uniquement intervenir ce principe un peu simpliste. Dans [Adda *et al.*, 1997], les questions de normalisation des textes des corpus utilisés pour la reconnaissance de la parole sont détaillées et quantifiées pour le français. Par exemple, le traitement de l'apostrophe pose question, et dépend de l'impact sur la taille du lexique de considérer l'apostrophe comme un séparateur ou non. En français, l'apostrophe est très fréquente, et pour ne pas augmenter de manière exponentielle la taille du lexique, l'apostrophe est considérée comme un séparateur de mots. Par exemple, le syntagme *l'oralité* sera traité comme deux mots séparés *l'* et *oralité*. C'est le cas de tous les petits mots comme *l'*, *j'*, *s'*, *t'*, etc... Le choix de normalisation n'est pas le même en français et en anglais, comme il est expliqué dans le chapitre « The use of lexica in Automatic Speech Recognition », écrit par Martine Adda-Decker et Lori Lamel [Van Eynde, F. and Gibbon, D., 2000]. En anglais, l'utilisation de l'apostrophe est restreinte, et pour cette raison, l'apostrophe n'est souvent pas considérée comme séparateur de mots. Les mots *I'll*, *you've* ou *he's* correspondent chacun à une entrée du lexique. Une pratique commune consiste également à développer systématiquement les formes abrégées. Par exemple *I'll* est réécrit *I will*, tout comme *I will*. Enfin, ces exemples très précis de normalisation, qui relèvent d'une pratique aguerrie des systèmes, peuvent être différents

selon l'application visée. Pour les systèmes de dialogue et de reconnaissance de la parole conversationnelle, les mots composés sont modélisés pour tenir compte plus facilement des variantes de prononciations propres à la parole spontanée.

Toutes les langues ne séparent pas les mots à l'écrit, c'est le cas de nombreuses langues asiatiques comme le chinois, le japonais et le thaï. Pour ces langues des algorithmes de segmentation en mots sont utilisés en pré- et post-traitement. En général il s'agit d'algorithmes qui partent d'un lexique déjà constitué mais des techniques automatiques font l'objet de recherches encore actuellement. La deuxième édition d'une évaluation sur la segmentation en mots en chinois, le « Second International Chinese Word Segmentation Bakeoff », a eu lieu en 2005. Il apparaît dans le compte rendu des résultats [Emerson, 2005] que la gestion des mots OOV est le principal problème malgré les améliorations par rapport aux résultats de la première évaluation en 2003. Néanmoins en reconnaissance de la parole, l'unité de mesure des performances pour le chinois mandarin est le caractère et non le mot.

3.6 Les sous-unités : une alternative ?

Une sous-unité désigne une unité lexicale plus petite que le mot. La recherche d'unités lexicales peut conduire à des décompositions de mots en sous-unités qui font sens, il s'agit alors de morphèmes au sens linguistique. On préférera le terme « morphème », pour désigner les sous-unités en général, qu'elles soient de vrais morphèmes ou pas. Ce terme employé dans [Creutz & Lagus, 2005] est un peu l'équivalent du terme « phone » utilisé pour désigner les unités acoustiques d'un système de reconnaissance. Dans la littérature du traitement automatique de la parole, ce terme est substitué au mot *phonème* qui a une acception précise en phonétique et en phonologie.

Voici des exemples de décompositions trouvés dans la littérature pour plusieurs langues :

- En allemand, le mot *Schulelternbeiratsmitglieder* peut être décomposé en *Schuleltern+beiratsmitglieder* puis en *Schul+ eltern+ beirats+ mitglieder* [Adda-Decker, 2003].
- En turc, la phrase *Isteklerimizi elde ettik dedi* peut être décomposée en *Istekler+ imizi el+ de etti+ k de+ di* [Arisoy & Saraclar, 2006].
- En arabe standard, des sous-unités peuvent être des préfixes, comme dans [Xiang *et al.*, 2006] avec entre autres *Al, bAl, fAl, kAl, ll, wAl, b, f, k, l, s, w*.

Le tableau 3.2 donne le nombre de mots de lexiques utilisés dans des systèmes au LIMSI-CNRS, et les taux de mots hors-vocabulaire (OOV) associés, pour deux langues bien dotées (anglais et français) et deux langues peu dotées (amharique et turc). Une taille de lexique *classique* en reconnaissance de la parole grand vocabulaire est de 65k mots, et des taux d'OOV raisonnables sont des taux inférieurs à 1%. Pour l'amharique et le turc, des taux d'OOV entre 6,5% et 7% sont obtenus sur des corpus de développement d'une quinzaine de milliers de mots avec des lexiques deux à trois fois plus grands que les

<i>Langue</i>	<i>Taille lexicale (mots)</i>	<i>OOV(%)</i>
<i>Anglais</i>	65k	0,6
<i>Français</i>	65k	1,2
<i>Amharique</i>	133k	6,9
<i>Turc</i>	250k	6,5

TAB. 3.2 – *Comparaison de taux de mots hors-vocabulaire (OOV) pour deux langues à morphologie particulièrement riche (amharique et turc), dont le caractère peu doté accentue les forts taux d’OOV obtenus, et deux langues qui ont une morphologie « moins riche », mais surtout qui sont des langues très bien dotées (anglais, français).*

lexiques utilisés pour l’anglais et le français. Ces différences sont dues principalement à la différence de morphologie entre ces deux types de langue, néanmoins le caractère peu doté les accentue également.

La figure 3.3 représente l’évolution de la taille de lexiques (« unique words »), en fonction du nombre de mots total de corpus de parole (« word tokens ») pour différentes langues, et deux types de parole différents, de la parole spontanée (« spontaneous ») et de la parole préparée (« planned »). Cette figure est tirée de l’étude de la modélisation non-supervisée de la morphologie avec application à la reconnaissance de la parole [Creutz *et al.*, 2007].

La figure montre que pour le finnois, l’estonien, le turc, et dans une moindre mesure l’arabe (dialecte égyptien), ont, en raison de leur morphologie particulièrement riche, des tailles de lexique qui augmentent très vite avec la taille des textes. De très grands lexiques sont nécessaires pour avoir une couverture lexicale correcte. Pour ces langues, l’utilisation d’unités plus petites que le mot paraît donc très intéressante, en terme de taux d’OOV et de problèmes de manque de données textuelles. Il est remarquable également que les courbes du dialecte arabe et de l’anglais spontanés présentent des évolutions plus petites que les courbes de parole préparée correspondantes, illustrant le fait que la parole spontanée est moins riche en vocabulaire.

Le principe de décomposer les mots pour les langues à tendance agglutinante, ou plus généralement à morphologie riche, nous a paru transposable aux langues peu dotées, même dans le cas où leur processus de génération des mots n’est pas majoritairement compositionnel. La recherche d’unités lexicales qui donnent des taux d’OOV raisonnables semble être l’une des premières étapes lors de l’élaboration d’un système de reconnaissance pour une langue peu dotée. Dans [Choueiter *et al.*, 2007], des sous-unités intermédiaires entre des phonèmes et des syllabes sont utilisées dans un système de reconnaissance pour générer les formes écrites et les prononciations de mots hors-vocabulaire. Outre la diminution des taux d’OOV, l’utilisation de sous-unités peut être intéressante pour l’indexation de documents et la recherche d’informations. Dans [Park *et al.*, 2005], de bons résultats de recherche d’information sont obtenus même lorsque ces recherches sont effectuées sur des transcriptions présentant des taux d’erreurs de reconnaissance

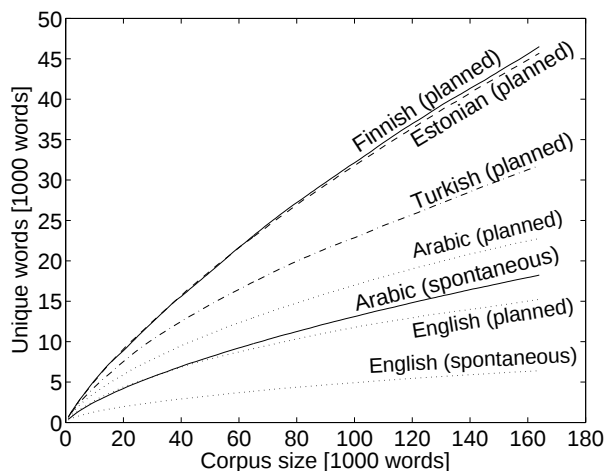


FIG. 3.3 – Évolution de la taille des lexiques en fonction du nombre de mots de corpus de parole pour l’anglais, le finnois, l’estonien, le turc et le dialecte arabe égyptien [Creutz et al., 2007].

très élevés, supérieurs à 35%. Le point important qui est souligné dans cette étude est la construction d’un lexique de reconnaissance qui donne les taux d’OOV les plus bas possibles, et qui contiennent les mots clés utilisés lors de la phase de recherche d’information. Dans [Nimaan et al., 2007] enfin, des sous-unités appelées racines sont utilisées pour la recherche d’information pour une langue peu dotée, la langue somali parlée notamment à Djibouti.

3.7 Segmentation morphologique

Deux grands types d’approches peuvent être différenciées : l’approche supervisée, qui fait appel à des connaissances linguistiques expertes et l’approche non-supervisée, avec peu ou pas de connaissances linguistiques.

Approches supervisées

Certaines langues bénéficient d’un outil appelé analyseur morphologique qui donne toutes les analyses morphologiques possibles d’un mot donné en entrée, c’est-à-dire que le mot est décomposé en sous-unités qui sont données avec leurs traits grammaticaux. Par exemple, un analyseur pour l’anglais pourrait donner pour l’entrée « books », deux sorties qui seraient « book + nom + pluriel » et « book + verbe + présent, 3^e personne du singulier », avec des balises à la place des mots entiers (par exemple « NN » pour nom commun,

« VB » pour verbe). Un exemple d'analyseur morphologique très utilisé est TreeTagger [Schmid, 1994]. Il a été adapté à plus d'une dizaine de langues, l'allemand, l'anglais, le français, l'italien, le néerlandais, l'espagnol, le portugais, le bulgare, le russe, le grec et le mandarin.

Pour construire un analyseur morphologique, les données nécessaires sont un lexique de mots « racines » ou lemmes, un lexique de morphèmes, les deux lexiques étant donnés avec les informations syntaxiques associées. Des listes de combinaisons possibles et impossibles de morphèmes (« propriétés morphotactiques ») ainsi que les règles associées (« propriétés morphographémiques ») sont également nécessaires. Par exemple en anglais, la combinaison « go+ed » est impossible. Un exemple de règle morphographémique consiste en un doublement de consonne comme lorsque « swim » est combiné à « ing » pour donner « swimming ».

Différentes implémentations sont possibles, il peut s'agir par exemple d'une base de données qui liste le plus grand nombre de mots possibles ou d'une approche par règles ou heuristiques (en arabe par exemple, il y a l'étude [Xiang *et al.*, 2006] qui sera décrite plus loin) ou d'approches par transducteurs à états finis (voir par exemple [Beesley & Karttunen, 2003]). Les méthodes par automates à états finis sont très efficaces et permettent aussi bien l'analyse que la génération morpho-syntaxique.

Approches non-supervisées

Ces méthodes font intervenir le moins possible de connaissances linguistiques. Elles peuvent éventuellement faire appel à des règles ou des heuristiques pour initialiser les modèles. Citons l'approche de Goldsmith (2001), fondée sur une approche de type Minimum Description Length (MDL), qui cherche à minimiser un coût de description de la morphologie d'un corpus de mots donné en entrée. Ce modèle tient compte de deux mesures, une mesure de concision (économie des symboles qui forment les lemmes, affixes et règles), et une mesure de la taille du modèle (une morphologie « moins bonne » donne une mesure plus grande). Goldsmith utilise une heuristique pour initialiser les hypothèses de décomposition des mots, une heuristique issue de l'algorithme de Harris sur la découverte de frontières de morphèmes [Harris, 1955].

L'étude [Adda-Decker, 2003] est un autre exemple d'utilisation de l'algorithme de Harris appliqué à l'allemand. Des gains de couverture lexicale sont constatés avec les règles de décomposition obtenues avec cet algorithme. Cet algorithme a été intégré dans le travail de cette thèse et il est décrit précisément dans le paragraphe 4.2.1 du chapitre 4 qui porte sur la sélection automatique des unités lexicales.

Un autre algorithme similaire à celui de Goldsmith est l'algorithme « Morfessor » [Creutz & Lagus, 2005]. À la différence de Goldsmith, aucune hypothèse sur le nombre de sous-unités qui composent les mots n'est faite, pour cette raison cette méthode ressemble aux méthodes de segmentation de textes. D'autre part son cadre purement probabiliste

correspond très bien au domaine de la reconnaissance de la parole et rend l'outil particulièrement simple à modifier. Pour ces raisons cet algorithme a été choisi comme base de travail pour mettre en place un paradigme de recherche automatique d'unités lexicales destinées à la reconnaissance de la parole. Cet algorithme sera décrit dans le prochain chapitre.

3.8 Différentes approches et résultats dans la littérature

De nombreuses études portent sur l'utilisation d'unités lexicales plus petites que le mot en traitement des langues. En traitement du langage naturel par exemple, l'utilisation des caractères à la place des mots est montrée comme une alternative tout à fait performante pour diverses applications (traduction automatique, génération d'énoncés) pour le japonais, qui ne possède pas de séparateur de mots [Denoual & Lepage, 2006].

En reconnaissance de la parole, l'utilisation d'unités lexicales plus petites que les mots n'est pas nouvelle mais fait l'objet de recherches actuelles en particulier pour des langues qui forment les mots par composition de morphèmes grammaticaux et/ou lexicaux. Parmi les études publiées figurent par exemple [Geutner, 1995; Kiecza *et al.*, 1999; Geutner *et al.*, 2000]. Des études plus récentes sont entre autres [Kirchhoff *et al.*, 2002; Adda-Decker, 2003; Xiang *et al.*, 2006; Kurimo *et al.*, 2006; Arisoy & Saraclar, 2006; Creutz *et al.*, 2007]. Ces différentes études mettent en oeuvre des approches différentes que nous allons décrire ci-dessous.

Une approche « Bottom-Up » est utilisée pour définir des unités de reconnaissance pour le coréen dans [Kiecza *et al.*, 1999]. Les mots coréens sont construits par concaténation de syllabes dont le nombre total est d'environ 3500 syllabes différentes. Utiliser directement les mots donnent des lexiques de très grande taille qui donnent des taux d'OOV très élevés, des taux de 30% sont reportés par les auteurs. Dans cette étude, les unités de reconnaissance de départ sont les syllabes. Les syllabes sont agglutinées de manière itérative en fonction des fréquences d'observation sur un corpus de textes, avec un critère d'arrêt sur le taux d'OOV (atteindre un taux d'OOV arbitraire de 5%) pour former des unités plus grandes que la syllabe, il s'agit d'unités polysyllabiques qui restent plus petites que les mots et qui permettent d'atteindre des taux d'OOV inférieurs à 1%. Les auteurs reportent des gains de performance en utilisant les unités plus grandes que les syllabes uniquement au niveau des phones et des syllabes. Au niveau des mots coréens, les systèmes construits avec les syllabes obtiennent de meilleurs résultats. Les auteurs ont remarqué que la plupart des erreurs sont faites sur les débuts et fins de phrase. Le nombre de syllabes étant plus grand que le nombre d'unités concaténées, le nombre d'erreur moyen est plus petit pour les systèmes basés sur les syllabes.

Les méthodes « Top-Down » sont plus représentées dans la littérature. Les méthodes « Bottom-Up » consistent à essayer de regrouper des graphèmes entre eux pour former des unités de taille comprise entre un caractère et un mot. Les méthodes « Top-

Down » tentent de faire l'inverse. Les mots sont décomposés en sous-unités, les plus petites unités étant les caractères eux-mêmes. Dans une application de reconnaissance de la parole, les sous-unités sont ensuite regroupées pour former des mots. Ces méthodes peuvent être séparées en deux groupes, les méthodes supervisées, basées sur des règles, et les méthodes automatiques non-supervisées, basées sur des corpus (en général un lexique avec les fréquences des mots). Les méthodes supervisées nécessitent des connaissances linguistiques de la langue étudiée. Elles consistent principalement à utiliser un analyseur morphologique sur les formes orthographiques des mots. Les méthodes non-supervisées tentent d'utiliser le moins possible de connaissances linguistiques, à l'aide principalement de propriétés basées sur les formes orthographiques des unités. L'intérêt de ces méthodes réside en général dans leur caractère indépendant de la langue. Elles peuvent être utilisées pour plusieurs langues avec un effort d'adaptation relativement faible.

Une fois que les mots ont été décomposés, le nouveau lexique composé de morphes peut être utilisé à différents niveaux de modélisation, la modélisation acoustique, la modélisation du langage, ou les deux. Dans la littérature, on trouve trois méthodes différentes que nous allons décrire en détail.

Rappelons tout d'abord l'équation 1.3, pour pouvoir cerner les avantages et les inconvénients des différentes méthodes. Dans cette expression, nous développons le terme $P(S|M)$ avec la somme des probabilités des prononciations H de la séquence de mots M . H correspond à une suite de modèles acoustiques.

$$\hat{M} = \operatorname{argmax}_M \sum_H P(H|M)P(S|H)P(M) \quad (3.1)$$

Une première méthode consiste à utiliser les morphes pour tous les éléments du système : les modèles acoustiques, le lexique et le modèle de langage. Parmi les études précitées figurent [Geutner, 1995; Geutner *et al.*, 1998]. Ces études portent sur l'allemand, mais imaginons que l'on travaille sur le français. Supposons que le mot *aller* ait été décomposé en *all-* *er*. Avec cette méthode seraient associés deux modèles acoustiques, un pour *all-* et l'autre pour *er*. L'avantage est de s'affranchir d'un lexique de prononciations, et on a dans ce cas $P(H|M) = P(M|M) = 1$. On imagine rapidement les limites de cette méthode, avec un inventaire de morphes dépendant de la langue étudiée, et également des modèles acoustiques plus complexes avec un nombre d'états supérieur à trois, qui dépendrait du nombre de phones par morphe, la nécessité de permettre des sauts d'états, etc. . .

Une deuxième méthode utilise les morphes dans le lexique et le modèle de langage, mais les unités acoustiques sont différentes (en général ce sont des phones). Les études concernées sont [Geutner *et al.*, 2000; Xiang *et al.*, 2006; Creutz *et al.*, 2007; Kurimo *et al.*, 2006]. À la différence de la première méthode, celle-ci ne fait pas l'économie d'un lexique de prononciations. Les nouveaux problèmes de cette méthode concernent en premier lieu le terme $P(H|M)$. En effet il faut être capable de définir les prononciations des sous-unités. Imaginons que le mot *apparaît* soit décomposé en *appara-* et *ît*. Quelle serait la

prononciation associée à *it* ? Dans ce cas précis, il faudrait avoir la possibilité d'avoir une prononciation vide, ce qui présente des difficultés lors du décodage. Une solution à ce problème consisterait à ne décomposer les mots uniquement si les sous-unités engendrées sont elles-mêmes des mots dont on est capable de donner une prononciation. Le terme $P(S|H)$ pose problème également, dans le sens où l'utilisation de petites unités accroît les confusions acoustico-phonétiques entre ces unités. Il faudra donc veiller à ne pas générer de sous-unités trop petites et trop semblables.

La troisième méthode utilise les morphes uniquement dans une phase de rescoring, c'est-à-dire qu'un modèle de langage fondé sur les morphes est utilisé pour rescorer des hypothèses issues d'un système basé sur les mots entiers. Les études concernées sont [Kirchhoff *et al.*, 2002; Arisoy & Saraclar, 2006]. Avec cette méthode, c'est uniquement le terme $P(M)$ qui est modifié. L'avantage de cette méthode est double, puisqu'elle s'affranchit du problème de confusion acoustique entre petites sous-unités, et du problème de définir des prononciations puisque ce sont les mots qui sont utilisés dans la partie acoustique du décodage. Plusieurs stratégies peuvent être envisagées, décomposer simplement les mots des N meilleures hypothèses et rescorer avec un ML de sous-unités, ou combiner les scores obtenus avec un ML de mots et un ML de sous-unités [Kirchhoff *et al.*, 2002], ou encore étendre les treillis d'hypothèses en ajoutant des arcs, avec des mots qui commencent avec un même préfixe par exemple [Arisoy & Saraclar, 2006].

Détaillons maintenant les résultats obtenus dans les différentes études mentionnées ci-dessus.

- *Méthode 1*

Dans [Geutner, 1995], qui porte sur l'allemand, les mots sont décomposés en morphèmes de deux manières, en utilisant des connaissances expertes de morphologie allemande, et d'une manière décrite comme « pas strictement linguistique » sans plus de précision. Les corpus d'apprentissage et de test sont issus du corpus de 250 dialogues « German Spontaneous Scheduling Task (GSST) » avec un total de 125k mots environ. Le lexique de départ contient 3,8k mots et les décompositions entraînent des réductions de 15% environ de taille de lexique. Comme précisé ci-dessus, les modèles acoustiques représentent les entrées lexicales elles-mêmes, c'est-à-dire les sous-unités issues des décompositions. Les modèles de langage sont également estimés sur les morphes. Des réductions de perplexité ont été observées, mais également une augmentation du taux d'erreurs de mots avec la décomposition strictement linguistique. En revanche avec le deuxième type de décomposition, les auteurs obtiennent un gain absolu de 0,7% sur un WER de référence de 35,3% avec un taux d'OOV de 9,0%. Il est précisé que cette seconde méthode décompose moins de mots que la méthode linguistique.

L'auteur décrit également une autre stratégie qui a consisté à garder les mots entiers pour apprendre les modèles acoustiques pour éviter la confusion acoustique des morphèmes due à leur petite taille. Les mots sont associés à leur forme racine pour procéder à une réestimation des scores dans les treillis d'hypothèses par un modèle de langage basé sur les

racines. Cette technique a conduit à une perte de performances par rapport aux modèles de mots entiers.

- *Méthode 2*

Dans [Geutner *et al.*, 2000], des expériences de décompositions en groupes de syllabes sont réalisées sur un système de reconnaissance du turc à grand vocabulaire (65k mots). Le corpus de parole contient au total 17 heures de parole lue (articles de politique et d'économie, journaux d'information) avec 100 locuteurs différents. Les performances sont comparées au niveau de ces groupes de syllabes et aucune amélioration n'est constatée par rapport au système de référence basé sur les mots entiers. Les auteurs mentionnent l'idée de confusion acoustique accrue due à la petite taille des sous-unités.

Dans [Creutz *et al.*, 2007], un algorithme appelé « Morfessor » est utilisé pour décomposer les mots dans un cadre probabiliste large, prenant en compte diverses propriétés comme la fréquence ou les séquences de caractères des mots par exemple. Les langues utilisées sont le finnois, l'estonien, le turc et le dialecte arabe égyptien. Tous les corpus d'apprentissage et de test sont des corpus de parole lue sauf pour l'égyptien, pour lequel le corpus est de type parole conversationnelle spontanée. Les modèles de langage (ML) et les modèles acoustiques de phones sont entraînés sur les corpus dont les mots ont été décomposés préalablement. Les modèles de langage sont d'ordre 4 pour le turc et le dialecte arabe. Pour le finnois et l'estonien où un algorithme dit de « n-gram growing » [Siivola *et al.*, 2007] augmente l'ordre des ngrammes tant que la vraisemblance du corpus d'apprentissage augmente. Pour ces deux langues, les ordres les plus grands sont respectivement 7 et 8. Des gains très significatifs sont obtenus avec les systèmes basés sur les morphes sauf pour l'égyptien pour lequel le système basé sur les mots est meilleur. Les résultats de cette étude sont résumés dans le tableau 3.3 avec les ordres des ML indiqués. La dernière ligne du tableau donne la différence absolue entre les WER mots entiers et morphes. Un delta positif donne le gain obtenu par le système basé sur les morphes. Un delta négatif signifie une perte de performances. Un gain de 7,6% absolu est obtenu pour le finnois avec un WER de référence de 17,8% avec des ML d'ordre variable contenant des 7-grammes pour le modèle de morphes et de mots. Pour l'estonien, un gain de 14,8% est observé avec un WER du système basé sur les mots de 48,1%. Pour le turc, les gains sont plus faibles, 1,2% avec des modèles de langage quadrigramme pour le système mots entiers (WER=32,6%) et le système basé sur les morphes. Avec un modèle trigramme pour le système mots entiers et 5-gramme pour le système morphes, le gain est de 5,5%. Le système morphes atteint un WER de 33,3% et est moins bon que les systèmes mots entiers et morphes construits avec un modèle de langage quadrigramme. Pour l'égyptien, le WER du système mots entiers est de 58,2% contre 59,9% pour le système morphes. Les auteurs n'avancent pas d'explication pour l'augmentation du taux d'erreurs uniquement observé sur le corpus de parole conversationnelle en dialecte égyptien.

Dans cet article sont étudiés également les taux de reconnaissance des mots qui étaient hors-vocabulaire avant l'application de Morfessor. Pour le turc par exemple, 13,2% sont

correctement reconnus avec le modèle quadrigramme. Les auteurs concluent que la représentation en morphes permet de reconnaître une partie des mots OOV sans détériorer la reconnaissance des mots du lexique.

	<i>Finnois</i>	<i>Estonien</i>	<i>Turc</i>	<i>Égyptien</i>
<i>Ordre ML mots</i>	Variable (7g)	Variable (8g)	4g	4g
<i>Ordre ML morphes</i>	Variable (7g)	Variable (8g)	4g	4g
<i>WER mots (%)</i>	17,8	48,1	32,6	58,2
<i>Delta (%)</i>	-7,6	-14,8	-1,2	+1,7

TAB. 3.3 – *Comparaison des performances des systèmes basés sur les mots entiers et sur les morphes pour le finlandais, l’estonien, le turc et le dialecte arabe égyptien. La dernière ligne donne la différence absolue entre le WER du système basé sur les morphes avec le WER du système basé sur les mots. Tous les scores sont donnés après recombinaison en mots entiers pour les systèmes basés sur les morphes. En bleu sont donnés les gains de performances et en rouge les pertes. Seul le système d’égyptien conversationnel présente une perte de performances.*

Comme dans [Creutz *et al.*, 2007], Arisoy et Saraclar (2006) utilisent l’algorithme Morfessor pour étudier la langue turque. Un taux d’erreurs de 36,0% est obtenu sur un corpus de parole de type Broadcast News avec 17 heures de données audio transcrites avec environ 250 locuteurs pour l’apprentissage. Pour le test, il s’agit d’une heure d’articles de journaux d’information lues par un unique locuteur, les articles sont différents de ceux qui ont servis pour l’apprentissage. Il faut remarquer que les modèles acoustiques modélisent directement les graphèmes, aucune conversion graphème-phonème n’a été utilisée. Le taux de mots OOV est de 5,6% avec un lexique de 120k mots. La décomposition non-supervisée a mené à un lexique de 34,3k mots avec un taux d’OOV nul. Les modèles acoustiques utilisés sont des triphones à états liés inter-mots (les diphtonges de fin de mot sont modélisés en triphones en tenant compte du phone de début du mot suivant). Les modèles comportent 5k états différents. Le taux d’erreur après recombinaison des morphes pour former les mots montre un gain relatif de 6% avec un WER initial de 36%.

Un dernier exemple d’étude qui utilise cette méthode est décrite dans [Xiang *et al.*, 2006]. C’est la seule étude mentionnée ici sur les décompositions d’unités lexicales pour une langue bien dotée, l’arabe standard, avec des tailles de corpus conséquentes et des systèmes à l’état de l’art pour la tâche de type broadcast news. Un jeu de 12 préfixes et 34 suffixes connus a été utilisé pour décomposer les mots d’un lexique arabe pour la tâche de reconnaissance de parole de type broadcast news. Le corpus audio d’apprentissage contient 150 heures de parole transcrite et le corpus de test contient environ 4 heures provenant de trois chaînes de télévisions différentes. Les modèles de langage ont été entraînés sur un total de 400 millions de mots. L’originalité de cette étude est de montrer que garder les N mots les plus fréquents intacts limite les erreurs introduites par les décompositions dues aux confusions acoustiques entre petites unités. Les mots les plus

fréquents sont bien représentés à la fois dans les transcriptions qui servent à estimer les modèles acoustiques mais également dans les textes qui servent à estimer les modèles de langage. La meilleure configuration de cette étude a été obtenue en gardant les 14k mots les plus fréquents intacts sur le lexique initial de 64k mots. Un gain absolu de 1,9% a été mesuré sur un score de référence de 19,3%.

- *Méthode 3*

[Kirchhoff *et al.*, 2002] est le rapport final d'un workshop de 2002 sur la recherche de nouveaux modèles de langage pour la reconnaissance de l'arabe moderne standard pour la parole de type conversationnel. Dans ce rapport, des techniques d'adaptation de l'arabe standard vers des dialectes sont également testées. Les trois principaux problèmes pour le traitement de l'arabe standard mentionnés dans le rapport sont le fait que la majorité des textes disponibles sont non-voyellés, l'importance de la variété lexicale et la complexité de sa morphologie. Dans les expériences décrites dans ce rapport, les sous-unités ne sont pas utilisées dans la modélisation acoustique mais uniquement dans la modélisation du langage, et seulement pour la phase de rescoring des N meilleures hypothèses de mots entiers (avec $N = 100$). Une expérience de décompositions en utilisant une simple liste d'une vingtaine de préfixes et suffixes connus a donné un gain absolu de 0,7% par rapport au score de référence de 55,1% de WER, alors que le modèle de langage basé sur les sous-unités montrait une perplexité normalisée plus grande que celle calculée avec un modèle de mots entiers. Ce gain a été obtenu en combinant les scores du ML fondé sur les mots et du modèle fondé sur les morphes.

Dans ce même travail, deux nouveaux modèles de langage ont été testés, l'un appelé « morphological stream model », qui ajoute des propriétés morpho-syntaxiques aux mots et l'autre appelé « factored language model » qui généralise les modèles précédents en permettant l'association de propriétés quelconques à chaque mot. Aucune expérience de reconnaissance n'a été effectuée avec ces deux types de modèles, néanmoins des mesures de perplexité montrent l'efficacité d'une technique de repli implémentée lors de ce workshop pour pouvoir utiliser les modèles de langage factorisés.

De toutes ces études il ressort que la modélisation morphologique peut apporter des gains importants de performances en reconnaissance de la parole. Pour les langues bien dotées, la possibilité des systèmes actuels d'utiliser des vocabulaires de très grande taille limite les problèmes posés par une morphologie riche. En revanche pour des langues peu dotées, l'augmentation de la couverture lexicale grâce aux décompositions est potentiellement intéressante.

La littérature du domaine est unanime quant à la confusion acoustique accrue entre les entrées lexicales issues des décompositions. Les morphes sont plus petits que les mots et dans certaines conditions une augmentation des taux d'erreurs de mots est observée. Dans les études [Kirchhoff *et al.*, 2002; Arisoy & Saraclar, 2006], les auteurs évitent le problème de la confusion acoustique entre petites unités lexicales et le problème de la définition des prononciations pour ces unités, en n'utilisant les décompositions des mots

uniquement dans une étape de post-traitement, en réestimant les treillis d'hypothèses ou les N meilleures phrases hypothèses. Cependant, les études plus récentes comme [Xiang *et al.*, 2006; Creutz *et al.*, 2007; Arisoy *et al.*, 2007] se servent des morphes au niveau du décodeur acoustique, pas seulement à l'étape de réestimation des scores des treillis, et ils obtiennent des gains tout à fait significatifs. L'idée de garder les mots les plus fréquents sans les décomposer [Xiang *et al.*, 2006] est également séduisante, et suggère d'utiliser un algorithme de décomposition qui tienne compte de la fréquence des mots. Dans le prochain chapitre, nous décrirons les méthodes que nous avons essayées pour limiter la confusion acoustico-phonétiques des petites unités.

Concernant la modélisation des différents éléments dans le système de reconnaissance, l'étude bibliographique exposée ci-dessus ne permet pas de déterminer avec certitude quelle approche est la meilleure. Les gains obtenus semblent dépendre de la langue traitée et du style de parole. La méthode 1 est d'emblée écartée puisque nous voulons travailler sur la transcription à grands vocabulaires.

Cependant les études les plus récentes citées ci-dessus utilisent la méthode 2, pour laquelle des modèles acoustiques de phones sont entraînés sur des transcriptions dont les mots ont été décomposés. La dépendance des modèles vis-à-vis de la position intra- et inter-mot devrait permettre d'obtenir des modèles de phones plus précis. Nous avons donc opté pour cette méthode qui combine l'utilisation des décompositions morphologiques pour les modèles de langage et les modèles acoustiques de phones.

3.9 Remarque sur la recombinaison des morphes

Comme il a été expliqué dans l'introduction de ce chapitre, l'unité lexicale de référence utilisée pour mesurer les performances d'un système de reconnaissance est en général le mot et la mesure est le taux d'erreurs de mots. Pour pouvoir comparer les performances de deux systèmes dont au moins un utilise une représentation en morphes, il est nécessaire de pouvoir recombinaison les morphes en mots avant de mesurer les taux d'erreurs.

En général, dans les études pré-citées, un signe (par exemple un '+') est accolé aux préfixes pour pouvoir recombinaison les morphes entre eux. Une autre possibilité similaire est d'accoler un signe aux suffixes. Un signe indépendant est ajouté à la fin d'un groupe de morphes pour signifier une frontière de mot, c'est le cas de [Arisoy & Saraclar, 2006] avec le signe « # ». Cette dernière solution a l'avantage de ne pas distinguer un morphe qui serait à la fois préfixe et mot. La taille du lexique est plus petite que celles des lexiques obtenus avec les deux autres manières de marquer les frontières de mots.

3.10 Conclusion

Dans ce chapitre nous avons introduit la problématique de la recherche de sous-unités lexicales pour une application de reconnaissance de la parole. Beaucoup d'études existent dans la littérature, sur des langues variées, principalement des langues qui ont un processus de génération des mots par composition ou agglutination de morphèmes lexicaux comme l'allemand, le turc ou le finnois. D'autres études portent sur des langues comme l'arabe standard, qui composent plutôt des morphèmes grammaticaux, qui sont en général de petits affixes. Deux points communs entre ces types de langues et les langues peu dotées sont les taux de mots hors-vocabulaire très élevés, dus en premier lieu à une morphologie éventuellement riche et en second lieu au manque de textes pour les langues peu dotées. Le second point concerne l'estimation des modèles de langage, qui est plus difficile avec ces langues. Derrière le problème du grand nombre de mots inconnus, se cache le problème du manque de « prédictibilité » des modèles de langage. Si la recherche d'unités plus petites que les mots permet de diminuer les taux d'OOV, les modèles de langage ne sont pas forcément améliorés.

Pour les langues peu dotées, le premier problème qui apparaît et que nous avons tenté de résoudre, est le nombre de mots hors-vocabulaire très élevé. L'application d'algorithmes de décomposition morphologique nous a semblé très prometteuse pour ces langues. La littérature est abondante sur le thème de l'apprentissage supervisé et non-supervisé de la morphologie d'une langue. Nous avons opté pour l'algorithme « Morfessor », car il s'inscrit dans un paradigme probabiliste simple et classique en traitement de la parole, à savoir un critère de maximisation de vraisemblance. En outre, cet algorithme est aisément modifiable pour intégrer des propriétés liées à la reconnaissance de la parole. La description de cet algorithme et des modifications que nous proposons sont l'objet du prochain chapitre.

Enfin, nous avons justifié le choix de la méthode appelée « méthode 2 » dans ce chapitre, concernant l'utilisation des formes décomposées des mots dans le décodeur. Elle consiste à utiliser les morphes lors de l'apprentissage des modèles de langage bien sûr, mais également éventuellement lors de l'apprentissage des modèles acoustiques de phones, pour les modèles acoustiques dépendants de la position intra- et inter-mot. Les deux passes du décodeur ainsi que l'extraction de la meilleure hypothèse des treillis sont réalisées avec des modèles de langage fondés sur une représentation en morphes.

Chapitre 4

Sélection automatique de sous-unités lexicales

Dans le chapitre 1, nous avons identifié les langues peu dotées comme des langues pour lesquelles les données qui sont le plus difficilement accessibles sont les textes, en raison d'une faible production écrite numérisée. Le but d'un système de reconnaissance est de transcrire une langue correctement, et cela dépend du thème de l'émission à transcrire, de la date également. Les données d'apprentissage doivent donc être, dans la mesure du possible, proches des données à transcrire. Pour les langues peu dotées, le principal problème est le nombre élevé de mots hors-vocabulaire (mots OOV pour Out-Of-Vocabulary), qui limite fortement les performances des systèmes de reconnaissance. En effet, comme il a été mesuré sur des langues bien dotées, un mot OOV est susceptible de générer 1,5 à 2 erreurs en moyenne. On peut imaginer aisément que cette proportion peut être plus élevée encore pour les langues peu dotées, en raison de la faiblesse plus importante des modèles de langage. La réduction des taux d'OOV a donc été l'un des axes majeurs de travail de cette thèse.

Les mêmes problèmes de taux d'OOV élevés et de n-grammes peu représentés sont rencontrés pour certaines langues bien-dotées également. Il s'agit de langues qui ont un processus de formation des mots par composition soit d'unités lexicales (appelées lexèmes ou morphèmes lexicaux) comme par exemple l'allemand, soit d'unités grammaticales (prépositions, articles possessifs, terminaisons verbales...) comme l'arabe standard par exemple. Par analogie avec le traitement de ces langues, nous avons essayé d'appliquer les techniques de découverte de morphèmes aux langues peu dotées.

L'une des manières de réduire le nombre de mots OOV consiste à décomposer les mots en plus petites unités que l'on peut regrouper pour former des mots. Dans ce chapitre nous exposons un paradigme statistique fondé sur une méthode « data-driven » ou d'apprentissage sur corpus, qui tente de trouver des décompositions de mots adaptées à la tâche de reconnaissance. Nous décrivons les apports et modifications réalisés sur une version

de base du programme.

4.1 Présentation de l'algorithme Morfessor

Morfessor [Creutz & Lagus, 2005] est un programme de traitement du langage naturel initialement, écrit en PERL, développé à l'université de technologies de Helsinki et distribué sous licence GNU/GPL. Le but de ce programme est de proposer une segmentation en morphèmes des mots d'un lexique. Les auteurs Mathias Creutz et Krista Lagus utilisent le terme « morphèmes » au sens linguistique usuel : unités élémentaires porteuses de sens. Néanmoins ils utilisent également le néologisme « morphe » pour désigner les sous-unités issues de la segmentation des mots par leur algorithme. C'est ce terme que j'utiliserai par la suite, en effet il n'est pas toujours possible de savoir si un découpage donne vraiment des unités porteuses de sens et le terme « morphe » permet de passer outre cette question.

Morfessor est un algorithme d'apprentissage non-supervisé qui ne nécessite aucune connaissance a priori de la langue étudiée et les morphes qu'il propose pour un lexique donné dépendent uniquement des mots de ce lexique. Morfessor propose deux modes d'utilisation possibles :

1. un mode « entraînement » : un modèle de découpage des mots d'un lexique donné (avec éventuellement avec les comptes d'occurrences des mots) est créé. L'entraînement est du type maximisation a posteriori (MAP) et utilise des propriétés exprimées sous forme de probabilités ou pseudo-probabilités comme par exemple la probabilité des séquences de caractères qui composent les mots du lexique.
2. un mode « décodage » : un modèle de décomposition des mots créé au préalable peut être utilisé pour découper un nouveau lexique de mots. Le choix des découpages de mots est réalisé à l'aide d'un algorithme de type Viterbi qui maximise les découpages donnant des unités les plus fréquentes possibles. En effet souvent plusieurs découpages sont possibles et dans ce cas, l'algorithme choisit en se basant sur la fréquence des morphes comme critère de sélection.

Nous avons travaillé sur la version 1,0 de Morfessor, qui date de 2005. Elle est téléchargeable à l'adresse suivante : <http://www.cis.hut.fi/projects/morpho/>

Voici un extrait d'un fichier créé par le programme de base sur un lexique de 65k mots français. Dans cet exemple, le lexique ne comportait pas de comptes de fréquences des mots, tous les mots ont une fréquence de 1.

```

1 évalua + it
1 évalua + nt
1 évalua + tion
1 évalua + tion + s
1 évalue
1 évalue + nt
1 évalue + r
1 évalue + ra
1 évalu + on + s
1 évalu + é
1 évalu + ée
1 évalu + ée + s
1 évalu + és

```

On voit dans cet exemple que les découpages proposés sont pertinents pour la plupart d'un point de vue sémantique, la marque du pluriel ou le suffixe « -tion » ont été isolés par exemple. L'algorithme peut proposer plusieurs découpages par mot, c'est le cas du mot « évalua + tion + s » par exemple. Notons la décomposition « évalua + it », dont nous avons parlé dans le chapitre précédent, qui poserait problème si nous avions besoin de donner une prononciation au suffixe *-it*, pour un lexique de reconnaissance.

4.1.1 Cadre mathématique de Morfessor

Dans ce paragraphe nous décrivons le mode « entraînement » du programme Morfessor.

L'algorithme cherche à maximiser de manière itérative la probabilité d'un lexique L en procédant à des découpages des mots, étant donné un corpus initial représenté par une liste de mots avec comptes d'occurrences (ces comptes peuvent être mis à 1 si l'on n'en dispose pas). Les mots sont traités un par un et lorsqu'un mot est décomposé en deux morphes, l'algorithme cherche à décomposer également les deux morphes résultants. L'équation 4.1 exprime le critère MAP utilisé, développé par la relation de Bayes :

$$\operatorname{argmax}_L P(L|\text{corpus}) = \operatorname{argmax}_L P(\text{corpus}|L)P(L) \quad (4.1)$$

où $P(\text{corpus}|L)$ est la *vraisemblance* du corpus étant donné le lexique L et $P(L)$ est la probabilité *a priori* du lexique L , i.e. la probabilité d'avoir M morphes distincts $m_1, \dots, m_M$.

Il est important de détailler ces deux termes pour pouvoir expliquer les apports et modifications que nous allons décrire plus tard, néanmoins pour tout besoin d'information complémentaire, l'article des auteurs [Creutz & Lagus, 2005] est très complet et très clair.

La vraisemblance du corpus $P(\text{corpus}|L)$

La différence entre ce que les auteurs de Morfessor appelle le corpus et le lexique est la suivante : le corpus correspond au lexique initial avec comptes éventuels pour chaque mot, le lexique lui, évolue en fonction des décompositions qui sont retenues. À chaque mot initial va être associée une forme finale décomposée ou non, et il est possible à partir de la sortie du programme de revenir sans ambiguïté à la forme initiale du mot. Le corpus correspond donc à la représentation des mots initiaux alors que le lexique correspond à la liste des morphes utilisés pour représenter le corpus. Par exemple si le lexique de départ comptait 10k mots et que le lexique après application de Morfessor ne comporte plus que 2k entrées lexicales, le « corpus » est toujours représenté par 10k entrées qui présentent les découpages associés, avec un total de 2k mots distincts.

Le terme $P(\text{corpus}|L)$ est explicité dans l'équation 4.2.

$$P(\text{corpus}|L) = \prod_{i=1}^N \prod_{j=1}^{n_i} P(m_{ij}) \quad (4.2)$$

Pour chaque $i^{\text{ème}}$ mot des N mots du corpus (en tokens), décomposé en n_i morphes, le produit des probabilités $P(m_{ij})$ de ces morphes est calculé, chaque morphe étant considéré comme indépendant des autres (aucune « grammaire » n'est utilisée). La probabilité pour un morphe m est estimée à l'aide de son nombre d'occurrences n_m comme indiqué dans la formule 4.3 où N est le nombre total de morphes en tokens.

$$P(m) = n_m/N \quad (4.3)$$

La probabilité *a priori* du lexique $P(L)$

le terme $P(L)$ de l'équation 4.1 fait intervenir des propriétés des mots qui constituent le lexique L :

$$P(L) = P(\text{propriétés}(m_1), \dots, \text{propriétés}(m_M)) \quad (4.4)$$

Dans la version de base de 2005, les propriétés retenues pour un morphe sont la fréquence du morphe dans le corpus et sa chaîne de caractères. La probabilité associée à une chaîne de caractères prend en compte également la taille de la chaîne. L'équation 4.4 s'écrit :

$$P(L) = P(n_{m_1}, \dots, n_{m_M})P(s_{m_1}, \dots, s_{m_M}) \quad (4.5)$$

où n représente la fréquence des morphes et s la chaîne de caractères. Cette dernière probabilité est décomposée en deux parties données dans l'équation 4.6, l'une estime la probabilité des séquences de caractères (notées s') et l'autre qui estime la probabilité de fin de chaîne (notée l).

$$P(s_{m_1}, \dots, s_{m_M}) = P(s'_{m_1}, \dots, s'_{m_M})P(l_{m_1}, \dots, l_{m_M}) \quad (4.6)$$

avec

$$P(s'_{m_1}, \dots, s'_{m_M}) = \prod_{i=1}^M \prod_{j=1}^{l_i} P(c_{ij}) \quad (4.7)$$

Pour Morfessor les auteurs ont fait le choix très simple de considérer les probabilités des suites de caractères comme les produits des probabilités de chaque caractère sans tenir compte des caractères voisins. En effet la probabilité $P(c_{ij})$ du $j^{\text{ème}}$ caractère du $i^{\text{ème}}$ morphe ne tient pas compte des caractères qui le précèdent. L'expression de $P(c_{ij})$ est simplement le rapport du nombre d'occurrences $n_{c_{ij}}$ du caractère dans tous les mots du corpus à la somme N_c des fréquences de tous les caractères présents dans le corpus 4.8.

$$P(c_{ij}) = n_{c_{ij}}/N_c \quad (4.8)$$

De même la probabilité des longueurs de morphes $P(l_{m_1}, \dots, l_{m_M})$ terme de droite de l'équation 4.6, s'exprime simplement par le produit des probabilités $P(l_{m_i})$ pour chaque morphe m_i , données par l'équation 4.9 :

$$P(l_{m_i}) = (1 - P(< /w >))^l P(< /w >) \quad (4.9)$$

où $< /w >$ est une balise de fin de mot. En pratique ils n'utilisent pas une balise de fin de mot mais ils ajoutent un espace à la fin des mots et assigne une probabilité à ce caractère de la même manière que pour tous les autres caractères.

L'équation 4.9 dit simplement que la probabilité que le morphe m_i soit de taille l est le produit d'avoir l caractères qui ne sont pas le caractère fin de mot, avec la probabilité du caractère de fin de mot. La probabilité de fin de mot est donc complètement indépendante de la chaîne de caractères qui le précède.

L'équation 4.4 fait également intervenir la fréquence des morphes. La valeur par défaut modélise la distribution de probabilités des fréquences de morphes indépendamment de la fréquence de chaque morphe. Cette probabilité ne dépend que du nombre total de

types M et de tokens N et a été choisie comme le nombre de manières de prendre N tokens dans une liste de M types. Les équations 4.10 et 4.11 donnent l'expression de cette probabilité.

$$P(f_{m_1}, \dots, f_{m_M}) = 1 / \binom{N-1}{M-1} \quad (4.10)$$

soit

$$P(f_{m_1}, \dots, f_{m_M}) = \frac{(M-1)!(N-M)!}{(N-1)!} \quad (4.11)$$

4.1.2 Décompositions d'un nouveau lexique à partir d'un modèle

Après avoir créé un modèle de découpage des mots d'un corpus, le programme Morfessor permet de charger ce modèle et propose des décompositions pour un nouveau lexique donné en argument. Plusieurs décompositions sont en général possibles pour une même unité lexicale et pour choisir la meilleure, un algorithme de type Viterbi est utilisé. L'algorithme propose les découpages de mots qui minimisent une fonction de coût qui porte uniquement sur le nombre d'occurrences des morphes résultant des différents découpages possibles. Cette fréquence est celle donnée dans l'équation 4.3.

Avec cet algorithme de type Viterbi, même des mots qui n'étaient pas dans le lexique qui a servi à créer le modèle de décompositions peuvent être décomposés. Dans cette phase de recherche de décompositions avec un modèle préalable, aucun nouvel apprentissage n'est réalisé.

4.2 Nouvelles propriétés et modifications apportées

Le paradigme de recherche de morphes de Morfessor est fondé uniquement sur des propriétés écrites ou graphémiques des unités lexicales. Aucune propriété à caractère oral n'est prise en compte. Nous avons déjà mentionné le phénomène de confusion acoustique, due aux découpages en sous-unités dont certaines ont une petite taille source de confusion (substitution par exemple), responsable de la diminution des performances de reconnaissance par rapport à des systèmes fondés sur les mots entiers. Pour pallier ce problème, l'idée a été d'introduire de nouvelles propriétés dans le paradigme de Morfessor qui orientent la stratégie de découpage des mots en morphes vers une application de reconnaissance vocale. Un premier paramètre utilisant les traits distinctifs des phones vise à favoriser les découpages de mots qui ont des traits distinctifs les plus distants possibles. Nous avons également ajouté la possibilité d'utiliser une autre expression de la probabilité de fin de mot en implémentant l'algorithme de Harris particulièrement efficace

pour un grand nombre de langues. Nous commencerons par détailler ce point. Des règles d'interdiction de décompositions ont également été introduites pour limiter la présence de morphes très proches phonétiquement. Des alignements graphémiques ont été réalisés pour déterminer quels étaient les graphèmes les plus susceptibles d'être confondus par le système. Ces contraintes seront décrites ci-dessous.

4.2.1 Algorithme inspiré de Harris

L'algorithme de Zellig Harris, inventé en 1955 [Harris, 1955], est un algorithme de détection de morphèmes qui initialement se fonde sur les représentations phonémiques des mots. L'idée générale vient de la constatation que pour les mots longs, un début de mot qui peut être suivi d'un nombre de caractères différents élevés est un bon candidat de morphème.

Dans [Adda-Decker & Adda, 2000; Adda-Decker, 2003], ce type d'algorithme a été appliqué directement sur les formes graphémiques des mots. Dans [Adda-Decker, 2003], des expériences de décomposition de mots avec cet algorithme ont été réalisées sur l'allemand, en vue d'une application à la reconnaissance de la parole. À partir d'un corpus de 300 millions de mots comportant 2,6 millions de mots distincts, des réductions d'OOV de 25% à 50% relatifs sont décrites. L'algorithme exploite le fait qu'un début de mot de k caractères a naturellement peu de caractères successeurs distincts possibles pour former des mots, qui existent dans la langue traitée, pour k suffisamment grand. Au rang $k+1$, ce nombre réduira davantage. Si ce nombre au contraire augmente pour un début de mot de k caractères, alors ce début de mot est un morphème candidat, pouvant se composer avec d'autres morphèmes commençant par des lettres distinctes variées. Ainsi l'algorithme compte le nombre de caractères successeurs distincts possibles pour tous les débuts de mots de taille k , et propose des frontières de morphèmes pour ces mots lorsqu'un maximum local est trouvé.

Cet algorithme est particulièrement simple, et il s'avère très efficace pour les langues qui ont un processus de formation des mots par composition de morphèmes. En allemand par exemple, la formation des mots est principalement fondée sur la composition de lexèmes, avec ajout parfois d'un élément de liaison appelé « Fungenelement », qui peuvent être les lettres *s*, *e* ou *n*. Par exemple le mot « Verkehrsabteilung » qui pourrait être traduit par « département des transports » est formé de « Verkehr » et de « Abteilung » qui signifient respectivement « transport » et « département ». La composition n'est pas le seul processus de formation des mots en allemand, les mots peuvent également être une nominalisation de syntagmes verbaux.

Le tableau 4.1 illustre le nombre de mots observés dans le corpus et le nombre de caractères distincts, qui peuvent compléter chacun des « préfixes » du morphème allemand « Verkehr ». Le *s* est le Fungenelement, qui marque ici une forme de génitif. Plus la taille du préfixe considéré augmente, plus le nombre de mots qui existent, et que l'on peut former avec ce préfixe diminue. Cependant, le tableau montre deux maxima locaux

k	préfixe	#mots	#successeurs
1	V	29k	24
2	Ve	17k	23
3	Ver	16k	28
4	Verk	1,7k	11
5	Verke	1,0k	6
6	Verkeh	0,99k	2
7	Verkehr	0,98k	12
8	Verkehrs	0,95k	29

TAB. 4.1 – Nombre de mots (*#mots*) et nombre de caractères distincts successeurs (*#successeurs*) pour les préfixes du morphème « Verkhers » calculés sur une liste de 1M de mots distincts

du nombre de successeurs pour les préfixes « Ver » et « Verkehrs ». Ainsi pour le mot « Verkehrsabteilung », l’algorithme de Harris proposera deux décompositions : « Verkehrsabteilung » et « Verkehrs abteilung ». Le choix entre ces deux formes dépend des motivations qui nous poussent à vouloir décomposer les mots, la deuxième forme étant peut-être plus satisfaisante si l’on s’intéresse aux lexèmes. La première forme est une application « bête » de l’algorithme, puisque « kehrsabteilung » n’a pas d’existence propre.

L’algorithme de Harris est également intéressant pour les langues qui ne composent pas forcément directement les lexèmes mais qui composent les morphèmes grammaticaux. Un exemple simple en français est la marque « s » de pluriel. De nombreuses langues agglutinent les articles possessifs, démonstratifs, pronoms, prépositions et postpositions. Dans des langues sémitiques comme l’arabe, l’amharique ou l’hébreu, les morphèmes grammaticaux sont effectivement collés aux mots qu’ils précisent. En arabe par exemple, il y a l’article « Al » qui est collé au mot qu’il précède.

Dans la version de base de Morfessor, la probabilité $P(l_m)$ d’avoir une frontière de morphème est donnée par l’équation 4.9. Cette probabilité est constante et indépendante de la chaîne de caractères qui le précède. Nous avons introduit la possibilité d’utiliser une autre probabilité qui n’est pas la même pour tous les mots et qui, inspirée par l’algorithme de Harris décrit ci-dessus, dépend de la chaîne de caractères du mot considéré. Cette probabilité est définie dans l’équation 4.12 comme la probabilité qu’un début de mot Pre soit un morphe (« Pre » pour signifier « préfixe »). Elle est égale au nombre de caractères différents $N_{succ}(Pre)$ qui peuvent compléter le préfixe Pre pour former des mots qui existent dans le lexique, divisé par le nombre total de lettres différentes N_c .

$$P_H(Pre) = N_{succ}(Pre)/N_c \quad (4.12)$$

$N_{succ}(Pre)$ est nécessairement grand pour les débuts de mots courts, cette définition

favorise les morphes courts ce qui paraît intéressant pour les langues qui composent des morphèmes grammaticaux qui sont très souvent des mots courts. Il faut remarquer que l'équation 4.12 ne correspond pas exactement à l'algorithme de Harris qui considère la variation absolue du nombre de caractères distincts potentiels successeurs.

4.2.2 Taille des morphes

Une option de taille minimale a été ajoutée pour les morphes générés par les décompositions. Les unités lexicales trop petites engendrent une confusion acoustico-phonétique accrue par rapport à des unités plus grandes.

Pour l'amharique par exemple, qui possède un syllabaire et non un alphabet, chaque syllabe est transcrite par au moins deux lettres. La taille des morphes en amharique est donc au minimum de deux caractères.

4.2.3 Propriété fondée sur les traits distinctifs

Les traits distinctifs au sens phonologique traditionnel correspondent à des propriétés phonologiques abstraites servant à discriminer des phonèmes entre eux. Selon Roman Jakobson [Jakobson *et al.*, 1952], les phonèmes d'une langue sont distinguables par un nombre limité de traits. Les informations véhiculées par les traits sont de type articulatoire comme le lieu d'articulation, le caractère voisé ou non-voisé, l'ouverture du conduit vocal par exemple et de type acoustico-perceptif comme le trait grave/aigu. Les systèmes de traits que l'on choisit pour décrire une langue ne sont pas forcément uniques et il n'y a pas d'accord unanime en général sur leur nombre et leur définition. Les traits que nous avons utilisés ici, dans le but d'introduire des propriétés donnant une notion de distance « acoustico-phonétique » lors de la sélection de sous-unités lexicales pour la reconnaissance, sont des traits binaires choisis à partir des descriptions phonétiques des langues traitées. Les traits choisis ne sont pas forcément suffisants si l'on voulait avoir une description de la langue. Ils recouvrent pour la plupart les classes de phones utilisées dans l'arbre de décision qui sert à partager les états des modèles acoustiques dans la phase d'apprentissage.

Exemples de traits distinctifs utilisés

Pour remplir les tableaux de traits distinctifs binaires, nous avons utilisé les méthodes de langues dont nous disposons qui donnent des détails sur la prononciation des sons mais également des tables que l'on trouve dans la littérature comme par exemple [Halle & Clements, 1983].

Le tableau 4.2 donne le sous-ensemble de traits distinctifs utilisés pour les voyelles turques. Les voyelles sont données avec leur forme graphémique, leur transcription pho-

nétiq ue choisie au LIMSI et le symbole de l'Alphabet Phon etique International correspondant. Les voyelles hautes sont les voyelles pour lesquelles la masse de la langue monte vers le palais.    l'inverse le corps de la langue est abaiss   pour les voyelles basses. Le trait arrondi correspond    la protrusion et l'arrondissement des l  vres ce qui   tr  cit l'ouverture de la bouche. Le trait r  duit d  signe la possibilit   d'omettre la voyelle. Seule la voyelle ı peut   tre r  duite, elle est assimil  e    un schwa. Le trait ant  rieur d  signe les voyelles qui sont prononc  es avec la masse de la langue d  plac  e vers l'avant de la bouche.

Trait	voyelles							
Graph��me	a	e	i	u	��	ı	o	��
Phone LIMSI	a	e	i	u	y	x	o	@
symbole API	a	e	i	u	y	ı	o	��
haut	0	0	1	1	1	0	0	0
bas	1	1	0	0	0	0	1	1
arrondi	0	0	0	1	1	0	1	1
r��duit	0	0	0	0	0	1	0	0
ant��rieur	0	1	1	0	1	0	0	0

TAB. 4.2 – *Sous-ensemble de traits distinctifs utilis  s pour les voyelles du turc*

Sont donn  s dans l'annexe 6.5 les traits distinctifs utilis  s pour les voyelles et les consonnes du turc ainsi que pour l'amharique. Les traits distinctifs utilis  s pour les consonnes sont : voisi   (les cordes vocales vibrent ou non), sonore (il n'y a pas de constriction qui bloque le passage de l'air. Le terme « sonante » est parfois utilis  ), glottalis   (la consonne est accompagn  e d'une fermeture de la glotte), coronal (la masse de la langue est d  plac  e vers les dents de la machoire sup  rieure), ant  rieur (la masse de la langue d  plac  e vers l'avant de la bouche), haut (la masse de la langue d  plac  e vers le palais), arri  re (la masse de la langue est r  tract  e), arrondi (il y a protrusion et arrondissement des l  vres), continu (pas d'occlusion du flux d'air dans la bouche), lat  ral (le placement de la langue emp   che un   coulement central de l'air dans la bouche), nasal (r  sonance des cavit  s nasales par abaissement du v  lum), strident (il y a pr  sence de bruit fricatif fort) et enfin affriqu   (la consonne est r  alis  e en deux phases, la premi  re phase est celle d'une occlusive et la seconde d'une fricative).

Int  gration    l'algorithme Morfessor

Les propri  t  s li  es aux traits distinctifs ont   t   int  gr  es    Morfessor en ajoutant un terme    l'  quation 4.5 qui devient l'  quation 4.13.

$$P(L) = P(f_{m_1}, \dots, f_{m_M})P(s_{m_1}, \dots, s_{m_M})D(td_{m_1}, \dots, td_{m_M}) \quad (4.13)$$

$D(td_{m_1}, \dots, td_{m_M})$ est le terme ajouté pour modéliser les traits distinctifs. Comme pour les deux autres termes, la propriété est considérée indépendante des autres morphes pour un morphe m_k donné, on a donc l'équation 4.14 :

$$D(td_{m_1}, \dots, td_{m_M}) = \prod_{i=1}^M D(m_i) \quad (4.14)$$

Le calcul de $D(m_k)$ a été restreint à la comparaison des traits distinctifs du morphe m_k avec les traits distinctifs des morphes qui ont même racine consonantale pour les traits des voyelles et même suite de voyelles pour les traits des consonnes. Par exemple pour les voyelles, l'expression de $D(m_k)$ est donnée dans 4.15 :

$$D_{td}(m_k) = \prod_{j=1}^{j=N_k-1} D_{td}(m_k, m_j) \quad (4.15)$$

avec

$$D_{td}(m_k, m_j) = \prod_{l=1}^{l=V_k} \frac{\Delta_{kl,jl}}{C} \quad (4.16)$$

N_k est le nombre de morphes qui partagent la même racine consonantique, $\Delta_{kl,jl}$ est le nombre de traits différents de la $l^{\text{ème}}$ voyelle des morphes m_k et m_j , V_k est le nombre de voyelles des morphes m_k et m_j et enfin C est le nombre total de traits différents considérés. Pour les voyelles turques par exemple, C est égal à cinq.

Pour les consonnes ce sont exactement les mêmes équations mais qui portent sur les suites de consonnes des morphes qui ont même suite de voyelles.

$D_{td}(m_k)$ a des valeurs comprises dans l'intervalle $[0, 1]$ et s'apparente à une distance et non à une probabilité puisque sur l'ensemble des morphes, D_{td} ne somme pas à 1.

4.2.4 Contraintes introduites dans la décomposition

La propriété liée aux traits distinctifs est théorique et ne prend pas en compte les variations phonologiques observées dans la parole réelle comme la prononciation des voyelles qui peut changer en fonction du contexte. En amharique par exemple, des alignements syllabo-tactiques ont permis de déterminer les confusions les plus fréquentes entre voyelles faites par le système (voir la section 2.6 pour plus de détails). Plus généralement pour chaque graphème d'une langue, les alignements aident à déterminer quels sont les phones associés qui sont les plus utilisés par le système d'alignement en fonction des modèles acoustiques de phones disponibles. Un lexique de prononciations associant tous les phones

élémentaires de la langue à chaque graphème est utilisé pour réaliser les alignements des données audio transcrites. Les paires de phones qui se substituent le plus souvent l'un à l'autre sont identifiées de cette manière. Elles apportent un moyen supplémentaire de prévenir la confusion phonétique engendrée par les décompositions des mots.

Nous avons restreint volontairement la détermination des paires de phones de confusion uniquement aux voyelles. Nous avons fait l'hypothèse que les consonnes sont moins sujettes à confusion les unes avec les autres que les voyelles.

Pendant la phase de création du modèle de décompositions, les morphes qui ne diffèrent des autres morphes que par un seul graphème sont comparés. Si la paire de phones associées à ces graphèmes fait partie des paires de confusion trouvées lors des alignements, la décomposition est interdite.

4.3 Implémentation de Morfessor

Dans cette section sont donnés quelques détails concernant l'implémentation de l'algorithme Morfessor et des nouvelles options.

4.3.1 Coût total

De manière classique en apprentissage statistique, les probabilités ne sont pas calculées telles quelles mais c'est l'opposé de leur logarithme (en base 2) qui est utilisé. L'utilisation du logarithme permet de sommer plutôt que de multiplier les probabilités qui peuvent être très petites parfois. Le produit de Bayes de l'équation 4.1 est donc remplacé par une somme qui peut être vue comme un coût à minimiser par les décompositions des mots.

L'équation générale 4.1 s'écrit comme une somme de deux coûts : un coût relatif au corpus qui est le terme de vraisemblance donné par l'équation 4.2 et un terme relatif au lexique qui correspond aux propriétés a priori, dont l'expression générale est donnée dans l'équation 4.5). Le terme relatif au lexique est lui-même composé d'une somme de coûts associés aux différentes propriétés : chaînes de caractères, longueurs, nombre d'occurrences et traits distinctifs des morphes. Les équations 4.17, 4.18 et 4.19 donnent l'expression du coût total :

$$\text{Coût total} = \text{coût corpus} + \text{coût lexique} \quad (4.17)$$

avec

$$\begin{aligned} \text{Coût lexique} = & \text{coût chaînes} + \text{coût longueurs} \\ & + \text{coût \#Occurrences} + \text{coût traits distinctifs} \end{aligned} \quad (4.18)$$

Soit

$$\begin{aligned} \text{Coût total} = & -\log(P(\text{corpus}|L)) - \log(P(f_{m_1}, \dots, f_{m_M})) \\ & - \log(P(s_{m_1}, \dots, s_{m_M})) - \log(D(td_{m_1}, \dots, td_{m_M})) \end{aligned} \quad (4.19)$$

À chaque itération, le programme cherche à décomposer les mots du lexique un par un en essayant toutes les décompositions possibles pour réduire le coût total. L'algorithme s'arrête lorsque la réduction du coût est inférieure à un seuil qui est fonction du nombre total de mots du lexique initial ou lorsque le nombre de morphes n'a pas diminué.

4.3.2 Probabilité de Harris

L'algorithme décrit dans la section 4.2.1 assigne une probabilité nulle de fin de mot à un préfixe si aucun caractère ne peut le compléter pour former un autre mot. En réalité on peut considérer qu'un mot est effectivement suivi par au moins un caractère qui est le caractère de séparation de mots. Dans l'implémentation de cet algorithme, nous avons donc décidé de ne pas prendre une probabilité nulle mais une probabilité minimale égale à l'inverse du nombre de caractères distincts. À cette probabilité minimale correspond un coût maximal qui peut être diminué par les décompositions. Lorsqu'une décomposition est validée par l'algorithme, nous avons choisi de ne considérer que le coût de Harris du préfixe. Le coût du suffixe associé n'est pas pris en compte pour pouvoir diminuer le coût lié à la taille des mots grâce aux décompositions. La raison en est que dans la plupart des cas de décompositions, le suffixe résultant de la décomposition a la même probabilité de Harris que le mot entier initial. Le coût de Harris du mot décomposé pourrait être la somme des coûts de Harris du préfixe et du suffixe mais dans ce cas, aucune diminution de coût serait observée. La partie « intéressante » de la décomposition est le préfixe dans le sens où il constitue une unité qui sert à former de nombreux débuts de mots du lexique.

4.3.3 Probabilité de nombre d'occurrences

Les coûts liés aux comptes ou nombre d'occurrences des morphes sont calculés suivant l'équation 4.11. Plus précisément, soient N le nombre de tokens ie le nombre d'occurrences des morphes et M le nombre de types c'est-à-dire le nombre de morphes distincts. Le coût de fréquences est calculé selon la formule 4.20, avec $\log(n!)$ approximé par $n \log(n-1)$. Les cas où le morphe est un singleton est distingué du cas où le morphe apparaît plusieurs fois.

$$\text{Coût} = \begin{cases} (M-1) \log(M-2) - (N-1) \log(N-2) & \text{si } (M-N) < 2 \\ (M-1) \log(M-2) - (N-1) \log(N-2) - (M-N) \log(M-N-1) & \text{sinon} \end{cases} \quad (4.20)$$

4.3.4 Exemple d'illustration

Pour illustrer l'algorithme et en particulier le paramètre des traits distinctifs, considérons un lexique de trois mots en mettant à 1 les nombres d'occurrences de ces mots pour faciliter l'interprétation. Ces trois mots font effectivement partie du lexique amharique. Le tableau 4.3 donne la liste de ces trois mots en script geez et en version transcrite avec un point pour figurer les frontières des caractères geez initiaux. Les décompositions dans cet exemple ne peuvent être placées qu'aux points qui séparent les caractères geez.

# Occurrences	Mot	Transcription
1	ነዋ	nE.wa
1	ነወ.	nE.wx
1	ነወ.	ni.wx

TAB. 4.3 – *Lexique de trois mots amhariques pris comme exemple pour illustrer le calcul des différents coûts. La première colonne donne le nombre d'occurrences, la deuxième et troisième colonnes donnent les mots en caractères ge'ez et en transcrit. Dans la forme transcrite un point a été ajouté pour figurer la séparation entre les deux syllabes qui composent chacun des trois mots.*

Morfessor dans sa version de base ne propose pas de décomposition pour ces trois mots. Morfessor modifié avec Harris non plus, à moins de donner un poids au coût relatif à Harris pour favoriser la décomposition. En revanche, si l'on ajoute le paramètre des traits distinctifs, le système propose la sortie suivante, sans avoir besoin de mettre de poids aux différents coûts :

```
1 nE + wa
1 nE + wx
1 niwx
```

Les deux premiers mots nE.wa et nE.wx ont été décomposés. Ils partagent la même première syllabe donc la décomposition paraît intéressante. Le troisième mot ni.wx est conservé tel quel.

Les coûts sont donnés dans le tableau 4.4, avant d'appliquer l'algorithme de décomposition (ligne « Initial ») et après décomposition (ligne « Final »). La ligne « Delta » donne les variations correspondantes soit la variation « Final - Initialisation ».

Le coût relatif à la vraisemblance du corpus, $-\log(P(\text{corpus}|L))$ de l'équation 4.19, est donné dans la colonne « Corpus ».

Les coûts associés aux propriétés du lexique décrites dans la section 4.1.1 dans le terme gauche de l'équation 4.19 sont détaillés : les coûts « chaîne » et « taille » associés aux

Coût	Total	Corpus	Lexique			
			chaîne	taille	fréquence	traits distinctifs
Initial	46,4	4,8	29,5	6,0	0,0	6,1
Final	44,7	9,6	24,9	3,0	3,3	3,9
Delta	-1,7	4,8	-4,6	-3,0	3,3	-2,2

TAB. 4.4 – *Les différents coûts avant décomposition (coût initial) et après décomposition (coût final) et les variations (delta) absolues correspondantes*

chaînes de caractères des mots du lexique, le coût « fréquence » et enfin le coût « traits distinctifs ».

Détaillons le calcul de chaque coût :

Coût de Corpus

Le nombre de morphes distincts est passé de 3 à 5, le coût de représentation augmente donc. C'est un effet dû à l'exemple pris ici, où les décompositions engendrent plus de morphes qu'il n'y en a dans le lexique initial. Ce n'est pas le cas lorsqu'on travaille avec un lexique de grand taille, les décompositions entraînant une réduction du nombre des unités lexicales.

Initialement nous avons :

$$\text{Coût}_i = -\log(P(\text{nEwa})) - \log(P(\text{nEwx})) - \log(P(\text{niwx})) \quad (4.21)$$

soit

$$\text{Coût}_i = -3 * \log(1/3) = 4,8 \quad (4.22)$$

Après décomposition :

$$\text{Coût}_f = -2 * \log(P(\text{nE})) - \log(P(\text{wx})) - \log(P(\text{wa})) - \log(P(\text{niwx})) \quad (4.23)$$

soit

$$\text{Coût}_f = -2 * \log(2/5) - \log(1/5) - \log(1/5) - \log(1/5) = 9,6 \quad (4.24)$$

Coût de chaîne de caractères

Avant décomposition, ce terme correspond à la probabilité des séquences de lettres qui constituent le lexique de trois mots. Le nombre de caractères distincts est égal à 12. Après décompositions, ce coût correspond aux probabilités des séquences « nE », « wa », « wx », « niwx ». La variation correspond donc à un gain égal à la probabilité de la séquence « nE ». Il y a au total 12 caractères utilisés pour les trois mots du lexique. La lettre « n » apparaît trois fois et la lettre « E » deux fois.

$$\begin{aligned}\Delta_{\text{chaîne}} &= -(-\log(P(\text{nE}))) \\ \Delta_{\text{chaîne}} &= \log(P(\text{n})) + \log(P(\text{E})) \\ \Delta_{\text{chaîne}} &= \log(3/12) + \log(2/12) = -4,6\end{aligned}\tag{4.25}$$

Coût de taille des morphes

Pour calculer les probabilités de fin de mot dans cet exemple, l'algorithme inspiré de celui de Harris, décrit dans la section 4.2.1, a été utilisé. Lorsqu'il n'y a pas de caractères qui peuvent compléter un préfixe pour former un mot, la probabilité est égale à l'inverse du nombre de caractères distincts (voir la section 4.3.2) soit $-3 * \log(1/4) = 6,0$ pour le lexique initial qui ne totalise que quatre caractères différents : **ɿ**, **ɿ̣**, **ɿ̥** et **ɿ̦**.

Après décomposition, le coût de Harris total correspond aux coûts des mots « nE + wx », « nE + wa » et « niwx ». Comme expliqué dans la section 4.3.2, les coûts des suffixes ne sont pas considérés. Le coût final est donc :

$$\begin{aligned}\text{Coût} &= -\log(P(\text{nE})) - \log(P(\text{niwx})) \\ \text{Coût} &= -\log(2/4) - \log(1/4) \\ \text{Coût} &= 3,0\end{aligned}\tag{4.26}$$

Le préfixe « nE » peut être complété par les syllabes « wa » et « wx » et sa décomposition a engendré un gain de 3,0 sur le coût total.

Coût de nombre d'occurrences

Le coût associé aux occurrences des morphes est nul initialement car les fréquences des mots ont été assignées à 1. Après décomposition, le morphe « nE » a une fréquence de 2, le nombre de tokens est égal à 5 et le nombre de morphes est égal à 4.

Le calcul du coût associé aux fréquences est réalisé à l'aide de l'équation 4.20 avec $M = 4$ et $N = 5$.

On a donc :

$$\begin{aligned}\text{Coût} &= -(3 \log(2) - 4 \log(3)) \\ \text{Coût} &= 3,3\end{aligned}\tag{4.27}$$

Coût des traits distinctifs

Dans cet exemple les morphes ne diffèrent que par les voyelles. Les coûts relatifs aux traits sont donc calculés sur les voyelles uniquement. À chaque mot sont associées la racine, constituée de la suite des consonnes et la suite de voyelles elle-même. Pour le mot « nEwa » par exemple, la racine est « nw » et la suite de voyelles est « Ea ». Les deux autres mots ont la même racine donc les traits distinctifs des suites de voyelles des trois mots du lexique vont être comparés. Les suites de voyelles à comparer sont « Ea », « Ex » et « ix ».

Pour le mot « nEwa », nous avons les comparaisons « E/E » et « E/i » pour la première voyelle « E » et la comparaison « a/x » comptée deux fois pour la seconde voyelle « a ».

Le score de ce mot vaut donc :

$$\text{Coût} = -1/2(\log(E/i) - \log(a/x) - \log(a/x))\tag{4.28}$$

Il y a un facteur 1/2 car comme le lexique est parcouru mot à mot lors de l'initialisation des coûts, les mêmes scores sont comptés deux fois.

Pour le mot « nEwx », le coût correspond aux coûts des paires de voyelles « E/i » et « x/a ».

Pour le mot « niwx », le coût correspond aux coûts des paires de voyelles « i/E » deux fois et « x/x ».

Le coût initial total vaut :

$$\begin{aligned}\text{Coût}_i &= 2,42 + 1,51 + 2,13 \\ \text{Coût}_i &= 6,1\end{aligned}\tag{4.29}$$

Voyons le coût final, après décompositions :

- « nE » : 0
- « wx » : « x/a »
- « wa » : « a/x »
- « niwx » : 0

Pour le mot « niwx », qui n'a pas été décomposé, le score lié aux traits est conservé, seuls les scores des mots décomposés sont modifiés. Ce choix peut être justifié par le fait que la décomposition d'un mot n'annule pas réellement la confusion au niveau des traits avec un mot non-décomposé pour lequel un score de traits distinctifs non-nul a été calculé

initialement. Dans notre exemple, la décomposition « nE + wx » ne change pas le fait que « niwx » est proche de « nE + wx » qui sera reconstitué en « nEwx » au final. En revanche il est clair que les morphes issus des décompositions ont des scores différents des mots dont ils faisaient partie initialement.

Le coût final total vaut :

$$\begin{aligned}\text{Coût}_f &= 1,81 + 2,13 \\ \text{Coût}_f &= 3,9\end{aligned}\tag{4.30}$$

Il y a une diminution de coût de 2,2 qui a permis les décompositions observées.

Coût total

Revenons maintenant au tableau général 4.4. Le delta total est négatif, il vaut $-1,7$, la décomposition des mots nEwa et nEwx a diminué le coût total. La somme des variations positives est de $4,8 + 3,3 = 8,1$, la somme des variations négatives donne $-4,6 - 3,0 - 2,2 = -9,8$. Sans la propriété des traits distinctifs, les variations négatives totaliseraient $-4,6 - 3,0 = -7,6$ ce qui est inférieur à $8,1$ en valeur absolue. Dans cet exemple, il n'y a donc pas de décomposition si le terme lié aux traits distinctifs n'est pas utilisé.

4.4 Conclusion

Dans ce chapitre, l'algorithme d'identification de frontières de morphèmes « Morfessor » a été présenté. Il consiste principalement à identifier de manière itérative quelles sont les décompositions de mots qui optimisent la représentation d'un lexique de mots avec comptes, en fonction de propriétés de nombre d'occurrences, de suites de caractères entre autres.

Les modifications et ajouts que nous avons réalisés sur l'algorithme de base sont orientés pour l'identification et la sélection d'unités qui sont destinées à être utilisées dans un lexique de système de reconnaissance de la parole. Les deux principaux ajouts sont l'interdiction de générer un morphe qui risque de se substituer à un autre morphe déjà présent dans le lexique. Le risque de substitution a été évalué sur la base d'alignements phonémiques préalables d'un corpus audio transcrit manuellement. Le deuxième apport est une propriété théorique basée sur les traits distinctifs qui donne une notion de distance acoustico-phonétique entre les morphes qui sont générés lors des décompositions. Sont favorisées les décompositions qui génèrent des morphes dont les phones présentent des traits les plus différents possibles des autres morphes. Nous avons restreint cette mesure aux seuls traits distinctifs des consonnes et des voyelles des morphes qui ont un nombre de phones identique.

Un exemple très simplifié d'un lexique de trois mots amhariques a été donné pour illustrer les calculs de chaque propriété. Il ressort entre autres que les modifications apportées permettent de décomposer plus d'unités que ne le fait l'algorithme de base.

Les prochains chapitres illustrent l'application de l'algorithme et le test de nos différents apports sur des systèmes de reconnaissance d'architecture standard.

Chapitre 5

Sélection automatique appliquée à l'amharique

Ce chapitre reporte et analyse les résultats obtenus par différents systèmes de reconnaissance pour tester l'approche de sélection automatique de sous-unités lexicales. Des systèmes de reconnaissance spécifiques à chaque représentation lexicale ont été construits et leurs performances sont comparées.

Les questions posées dans cette partie sont les suivantes :

- La diminution du nombre de mots hors-vocabulaire entraîne-t-elle de meilleures performances de reconnaissance ?
- Les mots qui étaient hors-vocabulaire et qui ne le sont plus dans les représentations avec mots décomposés sont-ils correctement reconnus ?
- Les décompositions génèrent-elles de nouvelles erreurs ?
- Les contraintes basées sur des alignements syllabo-tactiques préalables réduisent le nombre de décompositions, diminuent-elles réellement les erreurs entre morphes phonétiquement proches ?
- Les nouvelles propriétés liées aux traits distinctifs des phonèmes apportent-elles un gain significatif ?

5.1 Les différents systèmes comparés

Tous les systèmes ont la même architecture, comme tous les systèmes développés au cours de cette thèse. Deux passes de décodage sont réalisées avec une adaptation non-supervisée des modèles acoustiques après la première passe [Gauvain *et al.*, 2002]. Une ré-estimation des scores des treillis générés à la seconde passe est ensuite réalisée pour générer la meilleure hypothèse.

<i>Option</i>	<i>Signification</i>
BL	Système de référence (Baseline), pas de décomposition des mots
M	Morfessor (1,0)
M H	M + modification Harris
M H TDV	M H + traits distinctifs sur les voyelles
M H TDC	M H + traits distinctifs sur les consonnes
M H TDCV	M H + traits distinctifs sur les consonnes et les voyelles
Cc	+ contrainte de confusion déterminée par alignements graphémiques

TAB. 5.1 – *Tableau des abréviations des différentes options de décomposition.*

Plusieurs types de modèles acoustiques ont été testés : des modèles indépendants de la position inter- et intra-mot, appris sur les transcriptions en mots non-décomposés et des modèles acoustiques dépendants de la position inter- et intra-mot, spécifiques à chaque jeu d'options de découpage des mots. L'intérêt des modèles acoustiques indépendants de la position dans les mots est de pouvoir utiliser le même jeu de modèles pour toutes les représentations lexicales. Les modèles acoustiques dépendants nécessitent un apprentissage spécifique pour chaque représentation lexicale. Nous nous attendons à obtenir de meilleures performances avec les modèles spécifiques à chaque représentation lexicale qui sont plus précis que les modèles indépendants de la position a priori. Tous les jeux de modèles acoustiques couvrent environ 10,5k contextes différents avec un total autour de 8,5k états liés (32 gaussiennes par état). Le nombre de gaussiennes fixé à 32 est le même pour toutes les représentations lexicales testées.

Les différentes options de décomposition sont explicitées dans le tableau 5.1. Le système de référence, appelé BL pour baseline dans le tableau, est le système basé sur les mots, c'est le système de référence auquel on compare les performances des autres systèmes appris avec des unités lexicales décomposées. L'algorithme de base Morfessor est noté M dans le tableau et il correspond à l'algorithme utilisé sans modification (version 1,0). Les lignes suivantes du tableau correspondent aux options ajoutées une à une : l'option Harris expliquée à la section 4.2.1 et notée H, l'option des traits distinctifs notée TD et expliquée à la section 4.2.3 avec TDV pour les traits distinctifs uniquement sur les voyelles, TDC pour les traits distinctifs uniquement sur les consonnes et TDCV pour les traits distinctifs sur les voyelles et les consonnes. Enfin la contrainte décrite dans la section 4.2.4 visant à diminuer la confusion acoustico-phonétique entre les unités, a été testée sur tous les systèmes. Chaque système a été testé avec et sans cette contrainte.

Comme nous le verrons plus loin, l'algorithme inspiré de Harris permet de décomposer plus de mots que l'algorithme Morfessor standard. Nous n'avons donc pas testé les options supplémentaires sur l'algorithme sans l'option Harris, le but étant de décomposer le plus grand nombre de mots possible.

5.2 Les modèles de langage

Deux modèles de décomposition ont été appris pour chaque configuration, l'un sur les transcriptions et l'autre sur les textes Web. Le premier modèle a été créé avec le lexique avec comptes d'occurrences des transcriptions d'apprentissage des modèles acoustiques, tous les mots de ces transcriptions étant sélectionnés. Il totalise 50k mots distincts. Le second modèle a été appris sur le lexique issu des textes Web avec une sélection des mots en fonction de leur nombre d'occurrences : seuls les mots apparaissant au moins trois fois ont été inclus dans le lexique, totalisant 113k mots distincts.

Dans un deuxième temps, les modèles de décomposition ont été utilisés pour décomposer le lexique des transcriptions, identique à celui utilisé dans la première étape d'apprentissage, et pour décomposer le lexique des textes Web contenant cette fois tous les mots, sans limite minimale d'occurrence, soit 350k mots. La décomposition des mots se fait à l'aide du mode décodage de l'algorithme Morfessor, qui est un décodage de type Viterbi. L'algorithme trouve les décompositions de mots à partir d'un modèle existant, en tenant compte uniquement du nombre d'occurrences des sous-unités ou morphes produits.

Les décompositions obtenues avec l'algorithme Viterbi de Morfessor sur les transcriptions et les textes sont ensuite regroupées pour créer des règles de décomposition. Chaque configuration d'options a un ensemble unique de règles utilisées pour normaliser les textes et les transcriptions. De nouveaux modèles de langage et modèles acoustiques peuvent être appris à partir de ces transcriptions et textes dont les mots ont été décomposés.

Le décodage Viterbi sur le lexique des transcriptions ne donne pas les mêmes décompositions que l'apprentissage sur ce même lexique, il privilégie les morphes les plus fréquents. Enfin l'intérêt d'effectuer le décodage sur tous les mots issus des textes Web est de diminuer le nombre de mots hors-vocabulaire. Les mots n'apparaissant qu'une ou deux fois dans les textes peuvent être décomposés lors du décodage, même s'ils ne sont pas dans le lexique initial qui a servi à créer les modèles de décomposition.

Nombre d'entrées lexicales

Les contraintes de sélection des unités lexicales pour former les différents lexiques associés aux options de décomposition sont les mêmes que celles appliquées à la sélection du lexique initial de mots entiers. Tous les mots ou morphes des transcriptions ont été retenus, en revanche seuls les mots apparaissant au moins trois fois ont été sélectionnés pour les textes issus de l'Internet.

Le tableau 5.2 donne pour chaque jeu d'options le nombre d'entrées lexicales ou morphes retenus après décomposition. Pour pouvoir reconstituer les mots, un signe '+' a été accolé aux préfixes. La deuxième colonne du tableau donne le nombre de morphes du lexique réellement utilisé pour la reconnaissance. La troisième colonne donne, pour information, le nombre de morphes si l'on n'ajoute pas de signe '+' aux préfixes. La différence entre

<i>Options</i>	<i># Morphes '+'</i>	<i># Morphes</i>	<i>Réduction relative (%)</i>
BL	133384	133384	0,0
M	95937	70268	28,1
M Cc	128239	109694	3,9
MH	90740	65422	32,0
MH Cc	126105	107124	5,5
MH TDV	94198	69039	29,4
MH TDV Cc	128404	110321	3,7
MH TDC	66190	50062	50,4
MH TDC Cc	118596	101770	11,1
MH TDCV	66250	50192	50,3
MH TDCV Cc	107786	93572	19,2

TAB. 5.2 – *Nombre de morphes distincts et réduction relative des tailles de lexiques pour chaque jeu d'options de décompositions. BL : Baseline (mots entiers), M : Morfessor, Cc : contrainte de confusion, H : Harris, TDV : traits distinctifs des voyelles, TDC : traits distinctifs des consonnes, TDCV : traits distinctifs des consonnes et des voyelles.*

la deuxième et la troisième colonne correspond donc au nombre de morphes qui sont à la fois préfixe et mot. Enfin la dernière colonne donne les réductions de taille de lexique par rapport au lexique fondé sur les mots entiers qui sert de référence (noté BL pour Baseline).

Le lexique de départ comporte 133k mots. Les jeux d'options qui donnent les plus petits lexiques sont les jeux MHTDC et MHTDCV avec environ 66k morphes, ce qui correspond à une division par deux du nombre d'unités. Ces deux jeux se distinguent des autres par l'utilisation des traits distinctifs sur les consonnes. Comme nous l'avons vu dans le chapitre 4, l'ajout d'un nouveau paramètre dans l'algorithme diminue l'influence du coût qui porte sur les fréquences qui tend à empêcher les décompositions des mots fréquents. D'une manière générale, les traits distinctifs des consonnes sont plus nombreux que les traits des voyelles. Nous avons utilisé 14 traits contre 7 pour les voyelles. Nous avons constaté que les coûts associés aux consonnes sont environ d'un facteur dix plus grands que les coûts associés aux voyelles ce qui pourrait expliquer le plus grand nombre de décompositions lorsque les traits distinctifs des consonnes sont utilisés.

La contrainte Cc augmente le nombre d'unités d'environ 30%, un peu plus pour l'option MHTDC (44%). Le plus petit lexique avec cette contrainte est obtenu avec le jeu MHTDCV Cc avec 108k unités soit 19% de morphes en moins par rapport au lexique initial.

Taux d'OOV et vraisemblance

Le tableau 5.3 donne les taux de mots hors-vocabulaire (OOV) ainsi que les logarithmes des vraisemblances mesurés sur le corpus de développement qui contient 14,1k mots. Les taux d'OOV tokens tiennent compte de la fréquence des mots alors que les taux d'OOV types donnent la proportion de mots hors-vocabulaire distincts. Les log-vraisemblances ont été calculées en ne tenant pas compte des n-grammes qui contiennent un mot hors-vocabulaire. En effet les mots hors-vocabulaire sont remplacés par un mot balise (UNK) (pour « Unknown ») qui peut avoir une grande probabilité s'il y a beaucoup de mots OOV et cela peut biaiser la valeur de la vraisemblance.

Les modèles de langage (ML), construits avec la toolbox SRILM [Stolcke, 2002], sont des modèles quadri-grammes issus de l'interpolation de deux ML : l'un estimé sur les textes Web/journaux en ligne et l'autre sur les transcriptions manuelles des données d'apprentissage audio. Le coefficient d'interpolation a été optimisé en mesurant la perplexité sur le corpus de développement. Comme nous l'avons déjà signalé dans la section 2.3, pour des raisons de taille de corpus, le corpus de développement et de test sont un seul et même corpus ce qui revient à se placer dans des conditions un peu idéales. Pour chaque jeu d'options de décomposition, un ML spécifique a été estimé.

Les décompositions des mots entraînent une diminution « fictive » du taux de mots OOV due à l'augmentation artificielle du nombre d'unités par découpage, mais également une diminution véritable du nombre de mots qui étaient OOV au départ et qui ne le sont plus après décomposition. La réduction relative des taux d'OOV varie entre 30% et 45% selon les options par rapport au taux d'OOV de référence de 6,9%. Les plus bas taux d'OOV sont obtenus avec les options TDC et TDCV avec des valeurs respectives de 3,8% et 3,7%. Néanmoins les vraisemblances obtenues avec ces options sont plus faibles qu'avec d'autres options pour lesquelles le taux d'OOV est légèrement plus grand. Les taux d'OOV restent néanmoins bien supérieurs aux taux des langues bien dotées comme l'anglais américain ou le français, qui sont autour 1% et moins.

Les log-vraisemblances sont toutes plus petites que celle du ML de référence mots entiers BL, notamment à cause de l'augmentation du nombre d'unités et du fait que les sous-unités sont peu observées, même si elles sont plus fréquentes que les mots entiers. Cependant, si l'on compare les jeux d'options avec et sans confusion Cc, la vraisemblance est plus grande avec utilisation de la contrainte. La plus grande vraisemblance parmi les systèmes avec décomposition est celle du système MH TDV Cc. Nous allons constater dans la section suivante que ce système a été le plus performant.

Nous n'avons pas réalisé de mesures de perplexité sur les différentes représentations en préférant nous limiter à la vraisemblance qui n'a pas besoin de normalisation due aux différences de lexiques.

<i>Options</i>	<i>llh</i>	<i>OOV (%)</i>
BL	-30560	6,9
M Cc	-34785	4,3
	-34398	4,6
MH	-34845	4,2
MH Cc	-34422	4,5
MH TDV	-34795	4,2
MH TDV Cc	-34389	4,5
MH TDC	-35408	3,8
MH TDC Cc	-34653	4,4
MH TDCV	-35419	3,7
MH TDCV Cc	-34414	4,5

TAB. 5.3 – *Log-vraisemblances et taux d'OOV sur le corpus de développement/test. Ce corpus comporte 14,1k mots.*

5.3 Comparaison des performances des systèmes

Cette section décrit et analyse les expériences de reconnaissance menées avec les différentes représentations lexicales précédentes.

Des systèmes construits avec des modèles acoustiques indépendants et dépendants de la position inter- et intra-mot sont utilisés. Le système basé sur une représentation en mots entiers sert de référence, le taux d'erreurs de mots de ce système est de 24,0%. L'architecture et tous les paramètres des systèmes de reconnaissance sont identiques pour tous les systèmes.

Modèles acoustiques indépendants de la position intra-mot

Dans ce paragraphe sont comparées les performances de systèmes dont les modèles acoustiques sont indépendants de la position intra- et inter-mot (modèles IP pour « indépendants de la position »). L'avantage de ces modèles par rapport à des modèles dépendants de la position est de pouvoir utiliser un unique jeu de modèles pour tous les différents modèles de décompositions. Ce sont les lexiques de prononciation et les modèles de langage qui diffèrent d'un système à l'autre.

La taille des lexiques (colonne « # Morphes »), les taux d'OOV (Out Of Vocabulary) et les taux d'erreurs de mots (WER pour Word Error Rate) de tous les systèmes testés sont donnés dans le tableau 5.4. Les scores des systèmes ont été regroupés deux par deux : le score du système sans l'option de contrainte de confusion (Cc) et le score du système avec l'option. Tous les taux d'erreurs de mots donnés dans les tableaux sont les scores obtenus après recombinaison des sous-unités pour former des mots entiers.

<i>Options</i>	<i># Morphes</i>	<i>OOV</i>	<i>WER (%)</i>
BL	133384	6,9	26,2
M	95937	4,3	26,2
M Cc	128239	4,6	26,1
MH	90740	4,2	26,3
MH Cc	126105	4,5	25,7
MH TDV	94198	4,2	26,2
MH TDV Cc	128404	4,5	25,7
MH TDC	66190	3,8	26,6
MH TDC Cc	118596	4,4	26,2
MH TDCV	66250	3,7	26,6
MH TDCV Cc	107786	4,5	26,5

TAB. 5.4 – *Taille des lexiques en nombre de morphes, taux d’OOV et taux d’erreurs de mots (WER) pour tous les systèmes testés. Tous les systèmes utilisent les mêmes modèles acoustiques de phones indépendants de la position inter- et intra-mot (modèles IP).*

Les meilleures performances sont obtenues par les systèmes MH Cc et MH TDV Cc avec un taux d’erreurs de 25,7% et un gain de 0,5% absolu par rapport au système de référence. Le système M, Morfessor simple, est légèrement plus performant que le système MH qui découpe plus de mots. Les moins bons résultats sont obtenus avec les systèmes qui n’utilisent pas la contrainte Cc.

Modèles acoustiques dépendants de la position inter- et intra-mot

Dans ce paragraphe sont comparées les performances de systèmes dont les modèles acoustiques sont dépendants de la position inter- et intra-mot (modèles DP pour « dépendants de la position »), c’est-à-dire qu’un phone aura des modèles différents s’il se trouve en début, en milieu ou en fin de mot. Pour chaque jeu d’options de décomposition, des modèles acoustiques de phones spécifiques ont été appris. Nous avons tenté de conserver le même nombre de gaussiennes (32 gaussiennes en l’occurrence) pour tous les modèles. Le nombre de contextes modélisés varie un peu d’un jeu d’options à un autre, il reste proche de 10,5k contextes.

Les résultats sont donnés en taux d’erreurs de mots (WER) et taux d’erreurs de morphes (MER pour Morph Error Rate) dans le tableau 5.5. Les tailles des lexiques et les taux de mots hors-vocabulaire (OOV) sont également rappelés, les lexiques et modèles de langage sont les mêmes que ceux utilisés avec les modèles IP. Pour calculer les taux d’erreurs de morphes, comme plusieurs décompositions sont parfois possibles pour un même mot, les morphes des hypothèses ont été recombinaés pour former des mots entiers puis les règles de décomposition correspondantes au modèle ont été appliquées sur les hypothèses. Les

mêmes règles sont appliquées pour décomposer les mots des phrases de référence. De plus, les préfixes composés d'un unique symbole Ge'ez (soit deux caractères latins) n'ont pas été comptés dans l'estimation du taux d'erreurs de morphes. Ces petits morphes correspondent principalement à des prépositions et des articles.

Globalement les performances des systèmes utilisant des modèles acoustiques DP qui sont spécifiques à chaque modèle de décomposition des mots sont meilleures que celles obtenues avec des modèles IP, reportées dans le paragraphe précédant. Les différences sont d'environ 2% absolus. Il est intéressant de remarquer que le classement des systèmes DP est semblable à celui des systèmes IP. Le meilleur système est celui qui utilise les traits distinctifs des voyelles ainsi que la contrainte de confusion, il s'agit du système MHTDVCc.

Le système de référence ou baseline BL a un taux d'erreurs de 24,0% ce qui paraît relativement performant par rapport aux petites quantités de données d'apprentissage, il faut rappeler que les performances sont mesurées sur un corpus de développement qui sert de test. Ce corpus est très proche en date du corpus d'apprentissage, et les deux corpus comportent des locuteurs en commun.

Avec l'algorithme standard Morfessor noté M, une très légère dégradation de 0,1% est observée malgré la réduction du taux d'OOV de 2,6%. Avec l'option de Harris notée H, une dégradation plus grande est observée, égale à 0,5% absolu. Le nombre d'entrées lexicales est de 90740 contre 95937 pour l'option M. L'option de Harris permet de décomposer plus de mots que l'algorithme Morfessor standard. Une explication possible de la dégradation observée serait la confusion supplémentaire introduite lorsque des petites unités lexicales sont générées par les décompositions.

Cette hypothèse d'une confusion acoustico-phonétique accrue est confortée également par les performances obtenues avec l'option des traits distinctifs. Plus la taille du lexique diminue, c'est-à-dire plus il y a de morphes de petite taille introduits dans le lexique, plus la confusion entre les entrées lexicales est grande. Les options TDC et TDCV aboutissent à des lexiques de 66k mots soit la moitié de la taille du lexique initial et les performances obtenues sont les moins bonnes avec des taux de 24,8% et 24,9%.

Les systèmes M, MH et MHTD, qui n'utilisent pas la contrainte de confusion Cc ont des performances inférieures au système de référence. En revanche les trois systèmes avec Cc montrent de légers gains absolus compris entre 0,1% et 0,4% soit 0,5% et 2,0% relatifs. L'algorithme de Harris semble utile puisque de plus petits lexiques sont obtenus en l'utilisant et de légers gains sont observés. Concernant l'option des traits distinctifs TD, un gain absolu de 0,7% est obtenu entre les systèmes MHTD et MHTDVCc. La meilleure performance est obtenue avec le système MHTDVCc qui utilise l'option des traits distinctifs avec un gain de 0,4% absolu par rapport au système de référence.

Les taux d'erreurs de morphes sont globalement inférieurs aux taux d'erreurs de mots ce qui semble intéressant pour des applications de recherche d'information par exemple. En effet si les décompositions séparent principalement les articles, prépositions et autres

<i>Options</i>	<i># Morphes</i>	<i>OOV</i>	<i>WER (%)</i>	<i>MER (%)</i>
BL	133384	6,9	24,0	
M	95937	4,3	24,1	23,4
M Cc	128239	4,6	23,9	23,2
MH	90740	4,2	24,5	23,4
MH Cc	126105	4,5	23,7	23,2
MH TDV	94198	4,2	24,3	23,1
MH TDV Cc	128404	4,5	23,6	22,9
MH TDC	66190	3,8	24,8	23,4
MH TDC Cc	118596	4,4	24,2	23,3
MH TDCV	66250	3,7	24,9	23,2
MH TDCV Cc	107786	4,5	24,1	23,1

TAB. 5.5 – *Taux d'erreurs de mots (WER) et de morphes (MER) pour tous les systèmes testés. Les systèmes utilisent des modèles acoustiques de phones dépendants de la position intra- et inter-mot (modèles DP).*

préfixes de petite taille, les morphèmes lexicaux de taille plus grande sont mieux reconnus que les mots complets. La recherche de mots/morphes clés par exemple peut en être facilitée. Si l'on compare par exemple le WER du système de référence (24,0%) avec le MER du meilleur système MH TD Cc (22,9%), le gain relatif atteint 4,6%.

En comparant plus précisément les types d'erreurs faites par les systèmes, à savoir les pourcentages de substitution, d'élision et d'insertion de mots, il apparaît que tous les systèmes qui utilisent une représentation lexicale de morphes présentent des taux d'élisions plus grands et des taux d'insertions plus petits. Ainsi les pourcentages d'élisions et d'insertions valent 2,2% et 2,4% pour le système Baseline. Pour le système M, ces pourcentages sont respectivement 2,8% et 2,1%, pour le meilleur système MH TDV, ils valent 2,6% et 2,0%. Le total des élisions et insertions est donc le même pour le système de référence que pour le système MH TDV, c'est le taux de substitution qui est légèrement inférieur pour ce dernier système (19,0% contre 19,4%). Pour les moins bons systèmes, MHTDC et MHTDCV, les taux d'insertion sont identiques à celui du système BL en revanche, les taux d'élision et de substitution sont plus élevés (2,9% et 19,5% respectivement). Une explication pourrait être que pour ces systèmes qui comportent plus de petites unités lexicales que les autres, la composition de petits morphes pour former un mot entier est plus sujette à erreurs que la reconnaissance directe du mot entier ou de morphes de taille plus grande. D'autre part, le choix entre les formes préfixes ou mots pour les mots qui existent à la fois sous forme de mot et de préfixe dans le lexique favorisent pour la plupart la forme des préfixes qui sont plus fréquentes en général. Ainsi ces préfixes seront accolés au mot suivant et un plus grand nombre d'élisions sera observé. Les taux d'élisions et d'insertions sont équilibrés par des pénalités qui affectent les scores de vraisemblance. Nous avons gardé les mêmes pénalités pour les systèmes basés sur les

morphes que celles du système de référence basé sur les mots entiers, puisque la somme des taux d'insertion et d'élosion est constante.

L'outil « `sc_stats` » fourni par NIST [Pallett, 1990] permet de faire des tests statistiques pour comparer deux sorties de systèmes de reconnaissance. Quatre méthodes différentes sont classiquement utilisées pour voir si deux systèmes en compétition ont des performances significativement différentes : les erreurs sur les mots (Matched Pair Sentence Segment test, noté MP), les taux d'erreurs par locuteurs (deux mesures : le Sign Test ou Signed Paired Comparison, noté SI et le test du Wilcoxon Signed Rank, noté WI) et les erreurs à l'échelle des phrases (McNemar test, noté MN). Le critère MP est considéré comme le test le plus puissant car il considère des unités de taille variable qui favorise un grand nombre d'exemples sur lesquels estimer les différences de comportement (il s'agit des erreurs de reconnaissance) des deux systèmes à comparer. Les segments considérés avec cette méthode sont soit les phrases entières soit les groupes de mots délimités par deux mots correctement reconnus. La moyenne de la différence du nombre d'erreurs faites par les deux systèmes sur tous les segments délimités ainsi est comparée à une moyenne nulle, que l'on obtiendrait si les différences de sortie des systèmes étaient dues au hasard. Le test noté SI compare les taux d'erreurs de mots pour les différents locuteurs ou des différents tours de parole des fichiers de test. Si un système est meilleur sur une majorité de sous-ensembles du test, alors cela peut prouver la significativité des différences des sorties des deux systèmes, même si ces différences sont petites sur les sous-ensembles en question. Le test de Wilcoxon est identique au test SI, la seule différence étant la pondération par l'amplitude des différences observées sur chaque sous-ensemble de test. Enfin le test MN peut être vu comme un Sign Test au niveau des phrases. Les systèmes sont comparés par rapport au nombre de phrases entièrement correctes. Tous les tests font intervenir le degré de significativité qui est une probabilité de seuil. C'est la probabilité d'observer les différences entre les sorties des deux systèmes en supposant que les deux systèmes donnent des résultats identiques.

Le tableau 5.6 donne les résultats des tests pour chaque jeu d'options entre les systèmes avec et sans option de contrainte. La mention « same » est utilisée si aucune différence significative n'est trouvée. De manière habituelle pour ce genre de test, le niveau de confiance est pris à 95%. Si une différence significative est trouvée, alors le meilleur système est indiqué. Lorsque les tests sont positifs, c'est toujours le système avec l'option Cc qui a été identifié comme le meilleur, pour tous les jeux d'options étudiés. Cependant, certaines paires de systèmes ne montrent pas de différences statistiquement significatives, c'est le cas du système Morfessor simple (option M). Tous les autres jeux d'option donnent des gains significatifs pour le critère MP qui est le critère relatif au WER.

Pour le système MH TDV Cc, le degré de significativité (probabilité d'observer les différences entre les deux systèmes en supposant que les deux systèmes donnent des résultats identiques) est estimé à $p = 0,002$ ce qui paraît tout à fait correct, le seuil classique étant de 0,005.

Pour essayer de mieux comprendre ces résultats, une analyse des erreurs a été réalisée

<i>Options</i>	<i>MP</i>	<i>SI</i>	<i>WI</i>	<i>MN</i>
<i>M</i>	same	same	same	same
<i>MH</i>	Cc	same	Cc	Cc
<i>MHTDV</i>	Cc	Cc	Cc	same
<i>MHTDC</i>	Cc	Cc	Cc	same
<i>MHTDCV</i>	Cc	same	Cc	same

TAB. 5.6 – *Tests de significativité statistique des différences de performances entre les systèmes avec et sans l’option de contrainte Cc. Les acronymes sont MP pour Matched Pair Sentence Segment, qui donne les erreurs sur les mots, SI pour Signed Paired Comparison et WI pour Wilcoxon Signed Rank qui donnent les taux d’erreurs par locuteurs et MN pour McNemar qui donne les taux d’erreurs à l’échelle des phrases. Les tests donnent la réponse « same » lorsque les tests estiment que les différences des hypothèses des deux systèmes comparés sont dues au hasard, sinon les tests donnent le nom du meilleur système.*

pour déterminer entre autres combien de mots qui étaient OOV initialement et qui ont été décomposés ont été correctement reconnus.

Dans le système de référence, construit avec les mots, le taux d’OOV est de 6,9% soit 971 mots hors-vocabulaire sur 14,1k mots. L’algorithme de découpage des mots peut décomposer des mots qui ne sont pas dans le lexique, cela permet de diminuer le taux d’OOV. Après découpage, ont été considérés comme hors-vocabulaire les mots dont au moins un morphe issu du découpage n’appartient pas au lexique. Avec le meilleur système, le système MHTDCc, des 971 mots hors-vocabulaire initiaux restent 676 mots hors-vocabulaire, ce qui constitue une diminution faible du taux d’OOV (30% relatif). 295 mots ont été décomposés en sous-unités toutes dans le nouveau lexique. Sur ces 295 mots qui peuvent donc être reconnus correctement par le système MHTDCc, 122 ont été effectivement reconnus soit 41% des 295 mots qui ne sont plus OOV. 122 mots correspondent à un peu moins de 1% des 14k mots de test or le WER ne s’est amélioré que de 0,4% ce qui signifie que les découpages ont entraînés des erreurs qui n’étaient pas faites par le système de référence.

Le tableau 5.7 donne les nombres de mots qui étaient OOV avant décomposition et qui ne le sont plus après. La troisième colonne donne le nombre de ces mots correctement reconnus et le pourcentage absolu de gain correspondant. Le corpus de devtest initial comporte 971 mots hors-vocabulaire. Environ 310 mots ne sont plus hors-vocabulaire après décomposition, cela signifie que des décompositions en morphes présents dans le lexique ont été proposées par l’algorithme pour ces mots. Entre 122 et 137 mots parmi ces mots « ex-oov » sont correctement reconnus par les systèmes.

Les gains sur les mots hors-vocabulaires sont pratiquement identiques entre les différents systèmes avec cependant un gain légèrement supérieur pour les systèmes sans contrainte

<i>Options</i>	<i># Mots ex-OOV</i>	<i># Mots reconnus</i>	<i>Gain absolu (%)</i>
M	309	137	1,0
M C _c	296	129	0,9
MH	312	137	1,0
MH C _c	297	129	0,9
MHTDV	310	137	1,0
MHTDV C _c	295	122	0,9
MHTDC	316	141	1,0
MHTDC C _c	300	126	0,9
MHTDCV	316	135	1,0
MHTDCV C _c	294	127	0,9

TAB. 5.7 – Nombre de mots qui étaient OOV avant décomposition qui ont été correctement reconnus. Pour le système de référence, basé sur les mots entiers, le taux d'OOV s'élève à 6,9% soit 971 mots hors-vocabulaire au total. Environ un tiers des 971 mots OOV ne sont plus OOV après décomposition.

C_c d'environ 1,0% contre 0,9% pour les autres systèmes. La contre-partie de ces gains a été l'émergence de nouvelles erreurs sur les mots qui n'étaient pas hors-vocabulaire, ce qui a diminué les gains totaux globalement. Il semble que le contrainte basé sur les alignements permet de limiter l'introduction de nouvelles erreurs puisque les performances des systèmes qui l'utilisent sont meilleures que celles des systèmes qui ne l'utilisent pas. Les mots OOV qui ont été reconnus par les différents systèmes sont pour la plupart les mêmes à une dizaine de mots près. Par exemple, le système MHC_c a reconnu 19 mots ex-OOV qui ne figurent pas dans la liste de ceux reconnus par le système MHTDV C_c. À l'inverse ce dernier système a reconnu 13 mots ex-OOV que n'a pas reconnu le système MHC_c.

5.3.1 Expérience complémentaire avec un ML d'ordre plus élevé

Lorsque les mots sont décomposés en sous-unités, la portée d'un modèle *n-gramme* est plus faible à l'échelle d'une phrase qu'avec une représentation fondée sur les mots entiers. Pour cette raison, la plupart des études citées précédemment, utilisent des modèles de langage d'ordre plus élevé que les modèles *3-grammes* et *4-grammes* habituels, typiquement des modèles *5-grammes* et *6-grammes*.

Nous avons mesuré les perplexités obtenues avec des modèles *5-grammes*, et constaté qu'elles diffèrent très peu de celles obtenues avec les modèles *4-grammes* que nous avons utilisés jusque là.

Par exemple avec le modèle de décomposition MHTDV C_c, qui a donné les meilleures performances, un ML *5-grammes* estime à 465 la perplexité du corpus de textes devtest

	<i>MH TDV</i>		<i>MH TDV Cc</i>	
	<i>4-grammes</i>	<i>5-grammes</i>	<i>4-grammes</i>	<i>5-grammes</i>
# 1- <i>g</i>	94k	94k	128k	128k
# 2- <i>g</i>	2,4M	2,4M	2,6M	2,6M
# 3- <i>g</i>	4,4M	4,4M	4,2M	4,2M
# 4- <i>g</i>	4,7M	4,7M	4,4M	4,4M
# 5- <i>g</i>	0	4,5M	0	4,2M
<i>ppl</i>	535	363	468	465
<i>WER</i> (%)	24,3	24,3	23,6	23,4

TAB. 5.8 – Nombres de *n*-grammes, perplexités et *WER* pour les ML *4-grammes* et *5-grammes* des deux jeux d'option *MH TDV* et *MH TDV Cc*. Malgré la forte diminution de *ppl* observée entre les deux modèles pour l'option *MH TDV*, les *WER* sont identiques.

au lieu de 468 pour le ML *4-grammes*. Nous avons procédé à une réestimation de la meilleure hypothèse à partir des treillis générés avec le modèle *4-grammes* (procédure appelée « rescoring »), avec le modèle *5-grammes*. Le taux d'erreurs de mots obtenus est de 23,4%, soit un gain légèrement inférieur à 1% relatif, par rapport au rescoring avec le modèle *4-grammes*.

Nous avons procédé de même avec le jeu d'options *MH TDV*, identique au jeu précédent mis à part la contrainte *Cc* qui n'est pas utilisée ici. Le nombre de décompositions est beaucoup plus grand avec ce modèle qui n'est pas contraint, et l'on pourrait penser que l'usage d'un ML d'ordre plus grand serait plus bénéfique que pour le modèle *MH TDV Cc*. Les résultats ne vont cependant pas dans ce sens. En effet, bien que la perplexité diminue de 535 avec un ML *4-grammes* à 363 avec un ML *5-grammes*, les taux d'erreurs de mots obtenu après rescoring des treillis d'hypothèses sont identiques pour les deux MLs, atteignant 24,3%.

Le tableau 5.8 résume les caractéristiques des ML *4-grammes* et *5-grammes* pour les deux jeux d'option, en donnant le nombre de *n-grammes*, les perplexités mesurées sur le corpus de devtest ainsi que les taux d'erreurs de mots issus des rescoring.

Compte tenu des différences modestes observées entre les modèles *4-grammes* et *5-grammes*, nous avons préféré travailler avec les modèles *4-grammes* dans toutes les expériences menées au cours de cette thèse.

5.4 Conclusion

Dans ce chapitre ont été présentés des résultats expérimentaux quantifiés sur les performances de reconnaissance en fonction de différents critères de sélection automatique d'unités lexicales par décomposition des mots du lexique de reconnaissance. La langue

qui a servi de test est l'amharique qui est une langue qui dispose de peu de ressources numériques à ce jour.

Les modifications de l'algorithme de décomposition automatique, décrites dans le chapitre précédent, ont été testées en vue d'orienter le choix des propriétés de sélection des unités vers une application de traitement de la parole et plus précisément de reconnaissance de la parole. Des propriétés théoriques basées sur les traits distinctifs des phonèmes, une contrainte pratique d'interdiction de décomposer des mots en morphes qui sont trop proches acoustiquement –la notion de proximité phonético-acoustique ayant été précisée par des alignements syllabo-tactiques préalables–, une nouvelle définition de probabilité de fin de morphème qui a permis de découper plus de mots en ajoutant une dépendance vis-à-vis du contexte local des chaînes de caractères –algorithme inspiré de l'algorithme de Harris– ont été introduites. Enfin, une contrainte de taille minimale de morphe a été ajoutée pour autoriser des morphes de deux lettres latines minimum ce qui correspond à un caractère amharique.

Les résultats sont globalement concluants quant à l'introduction de ces propriétés. Le meilleur système est celui qui fait intervenir toutes les propriétés sus-mentionnées, en restreignant la propriété des traits distinctifs aux voyelles. Tous les systèmes qui utilisent la contrainte appelée « Cc » d'interdiction de décomposition des unités trop proches acoustiquement, présentent des gains par rapport aux systèmes qui ne l'utilisent pas. Si le meilleur système prend en compte la propriété basée sur les traits distinctifs des voyelles, les résultats présentés ne suffisent pas pour affirmer que cette propriété est indispensable pour obtenir un gain de performances. Même si des tests statistiques ont révélé des différences significatives de performances, les gains restent modestes. Cependant, il semble que la propriété de confusion acoustico-phonétique entre unités mène systématiquement à des gains.

Globalement, le taux de mots hors-vocabulaire a été diminué de 35% relatifs environ par les décompositions. Il reste néanmoins élevé par rapport au taux classique de 1% pour les langues bien dotées. Le gain de performances observé avec le meilleur système atteint 0,4% absolu par rapport au système de référence fondé sur les mots. Le gain sur les mots qui étaient hors-vocabulaire avant décompositions atteint 1,0% absolu, il y a donc de nouvelles erreurs qui sont introduites qui limitent le gain global à 0,4%. La sélection d'unités plus petites que les mots semble donc apporter un réel gain de performances sur les mots hors-vocabulaire.

De nouvelles erreurs sont introduites par les unités plus petites qui sont principalement des erreurs lorsque les mots sont reconstitués. Ces erreurs sont souvent des regroupements de deux ou plusieurs morphes qui sont en réalité des mots distincts dans la référence, ce qui expliquerait les taux d'élisions plus grands observés par rapport au système basé sur les mots entiers. En revanche, les taux d'erreurs de morphes sont plus petits que les taux d'erreurs après recombinaison des morphes pour former les mots entiers. Le meilleur système présente un taux d'erreurs de morphes inférieur de 1,1% absolus et 4,5% relatifs par rapport au système de mots entiers. Les représentations en morphes

sont intéressantes pour des applications d'indexation et de recherche de documents, où les petits mots fonctionnels ne sont pas forcément importants, et sont ajoutés à une « stop list », et également pour normaliser les formes infléchies des mots en lexèmes.

Chapitre 6

Sélection automatique appliquée au turc

La langue turque nous a semblé être un deuxième cas d'étude approprié pour tester la sélection automatique d'unités lexicales non-supervisée. Dans la littérature, le turc est souvent cité comme la langue à tendance compositionnelle ou « agglutinante » typique. La formation des mots est réalisée en ajoutant des mots qui font office d'affixes, et progressivement des mots très longs peuvent être créés. D'autre part, à l'instar de l'amharique, la relation graphème-phonème est relativement simple, ce qui a contribué également dans le choix de travailler sur cette langue.

Deux études différentes ont pu être réalisées, l'une sur un corpus de parole lue comprenant un peu moins de cinq heures transcrites manuellement et la seconde sur un corpus de parole de type broadcast news qui totalise 70 heures environ. Les expériences de cette dernière étude ont été menées en collaboration avec Ebru Arisoy et Murat Saraclar, du groupe de traitement de la parole « Busim » de l'université stambouliote Boğaziçi.

6.1 La langue turque, une langue peu dotée ?

La langue turque a une histoire riche et mouvementée. Son écriture à l'aide de l'alphabet latin est récente puisqu'elle date de la réforme linguistique de 1928, menée par le président Atatürk. Cette réforme remplaçait l'abjad arabe par l'alphabet latin auquel furent ajoutés deux diacritiques pour les deux affriquées ç [tʃ], ş [ʃ] ainsi que la lettre ğ et la voyelle i sans point, notée ı.

Le turcologue Louis Bazin, auteur de « La réforme linguistique en Turquie » dans *La réforme des langues* (1985), fait le portrait suivant de la situation linguistique sous l'empire ottoman :

« Dans l'État islamique théocratique et multinational qu'était l'Empire ottoman, soumis

à une acculturation arabe et persane intense dans ses classes dirigeantes – et spécialement dans la classe intellectuelle, comme celle des ulémas –, la langue écrite officielle et littéraire était envahie de termes arabes et persans, de plus en plus éloigné du parler turc vivant, et inaccessible à la masse populaire turque. »

En outre, l’abjad arabe transcrivait mal la langue turque dans la mesure où, par exemple, il ne permettait de noter que trois voyelles, alors que le turc comptait huit voyelles.

Jusqu’à très récemment, les études en reconnaissance de la parole en turc font état de ressources audio et textes très restreintes. Les corpus de parole sont le plus souvent des corpus de parole lue de quelques heures. La première étude sur un corpus de parole « broadcast news » date de 2007, le corpus audio totalisant 70 heures de parole transcrites manuellement [Arisoy *et al.*, 2007]. Le tableau 6.1 résume par ordre chronologique les études que l’on peut trouver dans la littérature concernant la reconnaissance automatique grand vocabulaire en turc. Dans ce tableau, figurent le type de parole des corpus utilisés, la référence de l’article, le nombre d’heures d’audio transcrites utilisées pour l’apprentissage des modèles acoustiques et le nombre de mots des textes pour estimer les modèles de langage dans la colonne « Train », la durée du corpus de test, le taux de mots hors-vocabulaire (OOV) et enfin les taux d’erreurs de mots (WER) des meilleurs systèmes de ces études. Remarquons que [Salor *et al.*, 2002] ne donne pas de taux d’erreurs de mots, uniquement un taux d’erreurs de phones (PER : Phone Error Rate).

L’étude la plus récente, [Arisoy *et al.*, 2007], reporte un WER de 39,6% obtenu par un système de référence basé sur les mots. Ce taux est un peu surprenant, il semble très élevé en dépit de la quantité conséquente d’heures d’audio transcrit (71h) et les presque 100 millions de mots des textes de modélisation du langage. Les auteurs mentionnent néanmoins que la quantité de parole étiquetée « clean speech » (ou condition acoustique *f0*), différenciée de la parole spontanée, téléphonique, ou avec de la musique, ne représente que 38% du corpus total. Les auteurs distinguent les WER obtenus sur les différentes sous-parties du corpus de test sélectionnées en fonction de leurs conditions acoustiques. Sur les parties de parole *f0*, le WER est de 26,3% soit 34% relatifs de différence par rapport au taux d’erreurs global. L’utilisation de sous-unités donne un WER global de 34,6% qui est leur meilleur performance et de 19,4% sur la partie *f0*. Les autres études citées dans le tableau concernent des corpus de parole lue et reportent des taux d’erreurs très élevés pour ce type de parole. Les deux études [Hacioglu *et al.*, 2003] et [Salor *et al.*, 2002] utilisent le corpus de parole lue de LDC que nous avons également utilisé. Nous verrons que nous avons obtenu des performances bien meilleures que celles de ces deux études, mais néanmoins bien inférieures aux performances de l’état de l’art en parole lue, obtenues pour des langues bien dotées.

L’étude la plus ancienne citée ici, [Çarki *et al.*, 2000], donne un résultat un peu surprenant, qui semble très bon par rapport aux quantités de données utilisées et surtout par rapport au taux d’OOV. Le corpus de parole totalise 13 heures de parole lue par une centaine de locuteurs et le corpus de textes contient 16 millions de mots. Le taux d’OOV est de

Type de parole	Référence	Train		Test	OOV (%)	WER (%)
		Audio	Textes			
Lue (Globalphone)	[Çarki <i>et al.</i> , 2000]	13h	16M	0h40	14,2	16,9
Lue (LDC)	[Salor <i>et al.</i> , 2002]	5h40	2M	1h	15,0	29,3
Lue (LDC)	[Hacioglu <i>et al.</i> , 2003]	5h40	2M	1h	15,2	43,0
Lue	[Erdogan <i>et al.</i> , 2005]	34h	30M	3h	18,2	32,5
BN	[Arisoy <i>et al.</i> , 2007]	71h	96M	2h30	9,3	34,6

TAB. 6.1 – Résumé des expériences trouvées dans la littérature sur la reconnaissance grand vocabulaire en langue turque. Les colonnes Train donnent le nombre d’heures audio transcrites et le nombre de mots (M pour million) utilisés pour construire les modèles acoustiques et linguistiques. Les meilleures performances de chaque étude sont données en taux d’erreurs de mots (WER) à l’exception de [Salor *et al.*, 2002] dont le taux de 29,3% est un taux d’erreurs de phones.

14,2% et le taux d’erreurs de mots de leur meilleur système est de 16,9%. En général, il est admis qu’un mot hors-vocabulaire cause entre 1,5 et 2 erreurs de reconnaissance, les résultats obtenus dans cette étude sont donc étonnants.

6.2 Les données audio et textes

Le corpus de parole lue

Concernant les données audio, nous avons utilisé le corpus de parole lue fourni par le Linguistic Data Consortium qui s’appelle « Turkish Microphone Speech V 1,0, Middle East Technical University, LDC2006S33 ». Ce corpus consiste en un enregistrement de 120 locuteurs différents, hommes et femmes, qui ont lu une quarantaine de phrases phonétiquement balancées choisies aléatoirement parmi 2462 phrases [Salor *et al.*, 2006].

Le tableau 6.2 résume les caractéristiques des données audio : le nombre d’heures, de locuteurs et de mots pour le corpus d’apprentissage et pour le corpus de développement. Le nombre total de mots des transcriptions atteint 43k mots. Nous avons isolé une sous-partie du corpus pour constituer un corpus de développement (appelé Devtest) qui a une taille d’environ 10% du corpus d’entraînement. Douze locuteurs ont été tirés au hasard dans le corpus, six hommes et six femmes. D’autre part il y avait des phrases en commun dans les corpus d’apprentissage et de devtest, ces phrases ont été enlevées de l’apprentissage. Le nombre total de mots a ainsi diminué de 43k mots à 36,5k mots. Le corpus d’apprentissage Train totalise 1773 phrases distinctes et 3549 phrases en tout pour une durée audio de 4h10min. La durée correspondante aux phrases qui ont été enlevées du corpus audio Train est de 42 minutes.

Les données textuelles autres que les transcriptions manuelles proviennent de sites de

	<i>Train</i>	<i>Devtest</i>
<i>Durée</i>	4h10	0h31
<i># locuteurs</i>	108	12
<i># mots</i>	32,3k	4,2k

TAB. 6.2 – Nombre d’heures, de locuteurs et de mots des sous-ensembles d’apprentissage (*Train*) et de développement/test (*Devtest*) du corpus de parole lue LDC2006S33

<i>Source</i>	<i># Mots</i>
<i>Bogaziçi</i>	11,6M
<i>AKŞAM</i>	710k
<i>TG</i>	500k
<i>VOA</i>	300k

TAB. 6.3 – Nombre de mots des textes d’apprentissage par source

journaux d’information en ligne. Le tableau 6.3 récapitule les nombres de mots des textes utilisés dans l’apprentissage des modèles de langage par source. Le texte le plus grand comprend plus de 11 millions de mots avec des textes très variés comme des articles d’information, des articles spécialisés (médecine, technologie...), mais également de la littérature. Ce corpus appelé Bogaziçi par la suite, nous a été gracieusement donné par le groupe de recherche sur le traitement de la parole de l’université stambouliote « Bogaziçi ».

Contrairement à l’amharique, aucune translittération ou transcription n’a été appliquée car les outils utilisés pour construire les modèles de langage supportent les quelques caractères à diacritiques du turc. Les caractères spécifiques au turc comme les lettres ı et ğ par exemple ont donc été conservées. L’encodage utilisé est un encodage UTF-8.

Le corpus de parole de type broadcast news

Nous n’avons pas eu accès au corpus de parole collecté et transcrit par l’équipe Busim d’Istanbul, les expériences de reconnaissance sur ce corpus ont été réalisées par eux-mêmes. Ce corpus, transcrit manuellement, contient 71 heures de parole enregistrée quotidiennement à partir de quatre chaînes de télévision et radios. Un corpus d’apprentissage de 68,6 heures et un corpus de 2,5 heures de test ont été séparés [Arisoy *et al.*, 2007].

6.3 Expériences sur le corpus de parole lue

Cette section décrit les expériences menées sur le corpus de parole lue uniquement.

Les taux d'erreurs très élevés obtenus sur ce corpus minimisent l'interprétation que l'on peut donner aux comparaisons entre les différentes options de décomposition. L'inadéquation du modèle de langage principalement estimé sur des textes d'information très éloignés d'un corpus de phrases créées spécifiquement pour être phonétiquement équilibrées, limite fortement les performances obtenues.

Le lexique de prononciation

Un lexique de 246k mots a été sélectionné à partir des transcriptions et des autres textes. Un seuil de fréquence minimale de 3 occurrences a été appliqué sur le texte Bogaziçi. Pour les autres textes, transcriptions, AKŞAM, TG et VOA, aucun seuil n'a été utilisé.

Dans la littérature, certaines études sur la reconnaissance du turc précisent ne pas utiliser de variantes de prononciation et utilisent directement les graphèmes [Arisoy & Saraclar, 2006; Arisoy *et al.*, 2007]. D'autres utilisent quelques règles de conversion, comme par exemple [Çarki *et al.*, 2000].

Pour générer les prononciations, nous avons utilisé les règles décrites dans [Çarki *et al.*, 2000]. D'une manière générale, à chaque graphème peut être associé un unique son, sans variante. Nous avons donc utilisé les mêmes symboles pour les graphèmes et pour les phonèmes. Il y a néanmoins quelques exceptions : les consonnes /ğ/, /k/, /r/ et /v/.

La lettre /ğ/ n'est pas prononcée entre deux voyelles, elle provoque un allongement de la voyelle précédente, qui n'a pas été modélisé compte tenu du peu de données d'apprentissage. Placée devant une voyelle postérieure, elle se prononce comme un [w], et devant une voyelle antérieure, comme un [j]. Devant e, i, ö et ü, /k/ est éventuellement palatisé : [k] devient [kj]. Les différentes prononciations de la consonne /r/ dépendent de la position au sein des mots. En début ou milieu de mot, elle se prononce légèrement roulée comme [r], en revanche en fin de mot, elle se prononce en général [ɾ]. À la consonne « v » ont été associés le phone [v] mais également le phone [w] devant la voyelle /a/. Environ 15% des occurrences de /v/ ont été alignés avec le phone [w]. Enfin la voyelle « ı » qui est un i sans point, a été associée à un schwa, noté « x ».

Un premier dictionnaire de prononciations, permettant les élisions éventuelles des schwas, et tenant compte des variantes de prononciation de /ğ/ et de /v/, a été utilisé pour apprendre un premier jeu de modèles acoustiques. Un second dictionnaire avec plus de variantes, liées à /r/ et /k/, a servi à apprendre un autre jeu de modèles acoustiques. Deux systèmes standards ont été construits avec ces deux dictionnaires et une dégradation des performances de plus de 1,0% absolu a été observée avec le deuxième dictionnaire. Avec le premier lexique, 3850 contextes ont été modélisés par 1577 états. Avec le deuxième dictionnaire, 4450 contextes ont été modélisés par 1840 états. Dans ce dernier cas, il se peut que le plus grand nombre de contextes à modéliser avec peu de données d'apprentissage soit responsable de la dégradation. Compte tenu de ce résultat, nous n'avons utilisé que le premier lexique de prononciations.

Les modèles de phones turques

L'inventaire des voyelles turques est très symétrique. Il comporte huit voyelles qui se répartissent en deux sous-ensembles selon le trait antérieur/postérieur. Le tableau 6.4 donne les huit voyelles avec leur symbole de l'Alphabet Phonétique International (API). Les différentes voyelles à l'intérieur d'un même mot sont soumises à une règle, dite règle d'harmonie des voyelles. Dans un même mot, il ne peut y avoir que des voyelles d'un seul groupe, soit des voyelles postérieures (a, ı, o, u), soit des voyelles antérieures (e, i, ö, ü). La voyelle des suffixes accolés à un radical sera donc modifiée en fonction des voyelles du radical. Par exemple la marque du pluriel est le suffixe -ler après une voyelle antérieure et -lar après une voyelle postérieure.

	Non-arrondies		Arrondies	
	Ouvertes	Fermées	Ouvertes	Fermées
Postérieures	a,[a]	ı,[ə]	o,[o]	u,[u]
Antérieures	e,[e]	i,[i]	ö,[ø]	ü,[y]

TAB. 6.4 – *Inventaire des voyelles turques données avec leur symbole [API].*

L'inventaire consonantique est donné dans le tableau 6.5 avec les graphèmes et les symboles API associés.

Labiales	b [b], p [p], f [f], m [m]
Dentales	d [d], t [t], s [s], z [z], n [n], l [l], r [r]
Palatales	c [dʒ], ç [tʃ], ş [ʃ], j [ʒ] [j], ğ [j] [w]
Vélaires	g [g], k [k], v [v] [w]
Uvulaire	h [h]

TAB. 6.5 – *Inventaire des consonnes turques, données avec leur symbole [API].*

Un jeu de 29 phones a été utilisé. Il comprend 8 phones pour les voyelles ainsi que 21 phones pour les consonnes. De manière habituelle, trois modèles supplémentaires ont été ajoutés pour modéliser les hésitations, les respirations et le silence.

Des modèles de triphones à états liés avec 32 gaussiennes par état, 3 états par modèle, dépendants et indépendants de la position intra-mot ont été construits. Avec un peu plus de 4h de données audio d'apprentissage, environ 4k contextes ont pu être modélisés avec un total d'environ 1,5k états.

La modélisation linguistique

Comme pour l'amharique, le modèle de langage utilisé pour le turc est issu de l'interpolation de deux modèles quadri-grammes avec repli et lissage de type Kneser-Ney modifié.

Pour l'amharique, un modèle a été construit avec les transcriptions d'apprentissage des modèles acoustiques et l'autre a été estimé sur les textes collectés sur Internet. Pour les expériences avec le corpus de parole lue, la quantité de transcriptions est si petite par rapport aux textes collectés sur Internet, que nous les avons simplement ajoutées aux textes. Comme les textes ont été récupérés sur une même période de temps, nous avons estimé un seul modèle de langage pour les sources AKŞAM, TG et VOA (voir le tableau 6.3). Un second modèle de langage a été estimé sur le corpus Bogaziçi de 11M de mots, et le modèle final a été le résultat de l'interpolation de ces deux modèles.

Les phrases de devtest ont été enlevées du corpus d'apprentissage audio. Avec un lexique ne comprenant que les transcriptions d'apprentissage, composé de seulement 7,7k mots, le taux de mots OOV mesuré sur le corpus de transcriptions devtest atteint 41%. Avec le lexique réellement utilisé, composé de 246k mots provenant des transcriptions d'apprentissage et des autres textes, le taux d'OOV est de 6,5%, ce qui est très comparable au taux d'OOV de départ sur l'amharique (qui est de 6,9%). La perplexité mesurée sur le corpus de devtest est très élevée puisqu'elle atteint 1526. Le modèle de langage est donc très peu adapté au corpus de devtest composé de phrases phonétiquement équilibrées. La perplexité du corpus devtest de l'amharique obtenu avec le modèle de langage de base était beaucoup moins élevée, avec une valeur de 372.

Expériences de reconnaissance

L'architecture des systèmes de reconnaissance est la même que celle des systèmes utilisés pour l'amharique [Gauvain *et al.*, 2002]. Il s'agit de systèmes en deux passes avec une adaptation non-supervisée des modèles acoustiques entre les deux passes. Après la deuxième passe, une ré-estimation des coûts sur les treillis est effectuée, avec un poids plus grand donné au score linguistique lors de cette étape.

Le corpus d'apprentissage des modèles acoustiques comprend seulement un peu plus de 4h de données transcrites. L'utilisation de modèles indépendants de la position intra- et inter-mot a donné des taux d'erreurs en mots supérieurs à 40%. Pour cette raison les expériences ont été menées uniquement avec des modèles dépendants de la position intra- et inter-mot qui donnent de meilleurs résultats. Les taux d'erreurs obtenus avec ces modèles sont encore très élevés par rapport aux scores habituels pour cette tâche puisqu'ils atteignent environ 35% de taux d'erreurs de mots. La figure 2.6 de la section 2.9, qui donnait l'évolution du taux d'erreurs en mots en fonction de la quantité de données transcrites en amharique pour une tâche de type émissions radio-télé diffusées, montre qu'à ce point de fonctionnement, un taux d'erreurs d'environ 47% est attendu. Ce score est nettement supérieur à 35%. Cette différence est due en majeure partie à la nature de la tâche elle-même. La transcription de parole lue est plus favorable que la parole de type broadcast news, en termes de conditions acoustiques d'enregistrement qui sont idéales : pas de bruit, ni de changement de conditions acoustiques.

Les différents jeux d'options des modèles de décomposition des mots sont les mêmes que

ceux utilisés pour l'amharique. Ils sont donnés dans le tableau 5.1 du chapitre 5. Comme dans les expériences menées sur l'amharique, les modèles de décomposition des mots ont été générés en deux étapes. Dans un premier temps, un modèle de décomposition est appris à partir du lexique initial de 197k mots issus du corpus de textes Bogaziçi, qui ne contient que les mots apparaissant au moins trois fois. Dans un deuxième temps, ce modèle est utilisé avec l'algorithme Viterbi sur trois lexiques différents : sur le lexique Bogaziçi de 197k mots, le lexique de 135k mots issus des autres sources (AKŞAM, TG, VOA) et sur le lexique de 7,7k issu des transcriptions servant à l'apprentissage des modèles acoustiques. Les trois modèles de décomposition obtenus sont ensuite regroupés en un seul modèle qui sert à décomposer les mots des différents corpus de textes avant estimation des modèles de langage. L'utilisation de l'algorithme Viterbi est intéressante car elle favorise les décompositions les plus fréquentes.

Le tableau 6.6 donne pour chaque jeu d'options le nombre d'entrées lexicales ou morphes retenus après décomposition. Comme pour l'amharique, un signe '+' a été accolé aux préfixes pour pouvoir reconstituer les mots. La deuxième colonne du tableau donne le nombre de morphes du lexique utilisé pour la reconnaissance. La troisième colonne donne la réduction relative de la taille des lexiques par rapport à la taille du lexique de mots entiers, qui sert de référence.

Le lexique initial comporte 246k mots. Comme pour l'amharique (voir la section 5.2), l'option TDC des traits distinctifs pour les consonnes conduit aux lexiques les plus petits, avec 77k mots environ pour les options MH TDC et MH TDCV, ce qui correspond à une division d'un facteur trois environ du nombre de mots initial.

L'option Cc augmente la taille des lexiques de 30k unités environ pour tous les systèmes, sauf ceux qui utilisent l'option TDC des traits distinctifs sur les consonnes pour lesquels les lexiques augmentent de 60k unités. Ces réductions de taille de lexique sont plus importantes que pour l'amharique, cela est dû, entre autres, au plus grand nombre d'unités lexicales au départ, environ 250k contre 133k pour l'amharique.

Le taux de mots hors-vocabulaire (OOV) avec la modélisation en mots entiers (BL pour baseline) est de 6,5%. Les différentes options de décomposition diminuent ce taux pour atteindre des valeurs comprises entre 1,3% et 2,0% sans la contrainte 'Cc'. Cette contrainte donne des taux d'OOV légèrement plus grands, compris entre 1,8% et 2,8%. Le tableau 6.7 donne les taux d'OOV ainsi que les vraisemblances obtenues avec les différents modèles de langage. On remarque que la contrainte Cc diminue les vraisemblances des systèmes qui utilisent la propriété des traits distinctifs (consonnes et/ou voyelles), ce qui n'était pas le cas des systèmes en amharique. Cela pourrait s'expliquer par le grand nombre de morphes et des décompositions des mots en morphes qui sont eux-mêmes peu fréquents.

Le tableau 6.8 donne les taux d'erreurs de mots (WER) pour les différents systèmes testés, en rappelant les tailles des lexiques en nombre de morphes, ainsi que les taux de mots OOV. Le WER du système basé sur les mots est de 35,9%, taux qui est très élevé pour ce type de tâche.

<i>Options</i>	<i># Morphes</i>	<i>Réduction relative (%)</i>
BL	246119	0,0
M	161217	34,5
M Cc	190522	22,6
M H	160662	34,7
M H Cc	189307	23,1
M H TDV	159613	35,1
M H TDV Cc	190649	22,5
M H TDC	77087	68,7
M H TDC Cc	139049	43,5
M H TDCV	76970	68,7
M H TDCV Cc	139207	43,4

TAB. 6.6 – *Nombre de morphes qui composent les lexiques pour chaque jeu d'options et réduction de taille des lexiques (en %) par rapport au baseline BL. BL : Baseline (mots entiers), M : Morfessor, Cc : contrainte de confusion, H : Harris, TDV : traits distinctifs des voyelles, TDC : traits distinctifs des consonnes, TDCV : traits distinctifs des consonnes et des voyelles.*

Remarquons que les scores obtenus par les quatre systèmes qui utilisent les traits distinctifs sur les consonnes ne semblent pas cohérents avec les résultats obtenus sur la langue amharique, dans la mesure où l'option Cc donne une augmentation du WER. Le système M H TDC, par exemple, présente un WER de 35,2% contre 36,3% pour le même système avec l'option de contrainte en plus, le système noté M H TDC Cc.

Si l'on met de côté ces quatre systèmes, il apparaît que tous les systèmes basés sur les morphes, quelles que soient les options de modélisation, donnent un WER plus petit que le Baseline. Le meilleur système est le Morfessor baseline combiné avec la contrainte basée sur les alignements phonotactiques, noté M Cc. Le gain est de 5% relatifs par rapport au baseline. Le tableau 6.9 donne les taux d'erreurs de lettres (LER), qui correspond à un score de phones en raison de la conversion graphème-phonème directe pour cette langue. Ce tableau est très intéressant dans la mesure où il montre que tous les LER des systèmes de morphes sont inférieurs au LER du système fondé sur les mots entiers. Cela signifie que de nouvelles erreurs sont introduites lors de la recombinaison des sous-unités en mots. Ces erreurs sont en particulier plus nombreuses pour les systèmes qui utilisent l'option TDC qui présente le plus grand nombre de décompositions.

La différence de comportement des systèmes qui utilisent la propriété TDC est peut-être due à de trop nombreuses décompositions de mots. En effet les ratios du nombre de morphes divisé par le nombre de mots au départ, calculés sur les textes d'apprentissage des modèles de langage, sont bien plus élevés pour les quatre systèmes qui utilisent la propriété TDC. Le tableau 6.10 donne les valeurs de ce ratio pour chaque système. Il varie entre 1,10 et 1,20 pour ces systèmes et entre 1,02 et 1,04 pour tous les autres systèmes.

<i>Options</i>	<i>llh</i>	<i>OOV (%)</i>
BL	-12031	6,5
M	-13366	2,0
M C _c	-13355	2,4
MH	-13369	2,0
MH C _c	-13364	2,3
MH TDV	-13371	2,0
MH TDV C _c	-13380	2,5
MH TDC	-13291	1,3
MH TDC C _c	-13354	1,8
MH TDCV	-13294	1,3
MH TDCV C _c	-13354	1,8

TAB. 6.7 – Log-vraisemblances et taux d’OOV sur le corpus de devtest. Le corpus de devtest a 3,3k mots et 2k mots distincts.

<i>Options</i>	<i># Morphes</i>	<i>OOV (%)</i>	<i>WER (%)</i>
BL	246119	6,5	35,9
M	161217	3,9	35,0
M C _c	190522	2,4	34,1
MH	160662	2,0	35,4
MH C _c	189307	2,3	34,5
MH TDV	159613	2,0	35,5
MH TDV C _c	190649	2,3	34,7
MH TDC	77087	1,2	35,2
MH TDC C _c	139049	1,8	36,3
MH TDCV	76970	1,2	34,7
MH TDCV C _c	139207	1,8	35,8

TAB. 6.8 – Nombre de morphes des lexiques, taux d’OOV, taux d’erreurs de mots (WER) des différents systèmes testés sur le corpus de parole lue. NB : tous les WER sont donnés après recombinaison des sous-unités pour former des mots entiers

<i>Options</i>	<i>LER (%)</i>
BL	14,6
M	12,4
M C _c	12,3
MH	12,9
MH C _c	12,5
MH TDV	13,0
MH TDV C _c	12,6
MH TDC	12,7
MH TDC C _c	12,9
MH TDCV	12,5
MH TDCV C _c	12,7

TAB. 6.9 – Taux d’erreurs de lettres des différents systèmes (LER) testés sur le corpus de parole lue.

<i>Options</i>	<i># Morphes / # Mots</i>
BL	1,00
M	1,04
M C _c	1,02
MH	1,04
MH C _c	1,02
MH TDV	1,04
MH TDV C _c	1,02
MH TDC	1,20
MH TDC C _c	1,10
MH TDCV	1,20
MH TDCV C _c	1,10

TAB. 6.10 – Ratios du nombre de morphes sur le nombre de mots calculés sur les textes d’apprentissage des modèles de langage.

Les résultats obtenus sur le corpus de parole lue en langue turque semblent montrer qu'il existe une valeur seuil du ratio du nombre de morphes divisé par le nombre de mots initial dans les corpus de textes. Cette valeur seuil avait été rencontrée dans les expériences menées sur l'amharique, puisque les systèmes qui décomposaient le plus de mots montraient de légères dégradations de performances. Des gains plus importants et systématiques avaient été observés avec l'usage de la contrainte sur les alignements phonotactiques préalables ; les expériences sur le corpus de parole lue en turc montrent que cette contrainte apporte un léger gain avec les systèmes M, MH et MHTDV mais une dégradation pour les systèmes MHTDC et MHTDCV. Ce résultat suggère de revoir la contrainte de confusion telle qu'elle est appliquée pour ce corpus, elle ne joue pas son rôle pour des modèles où le nombre de décomposition est plus grand, avec un ratio autour de 1,2.

Les résultats obtenus sur le corpus de parole lue sont néanmoins à relativiser compte tenu de la très faible quantité de données et surtout de l'inadéquation du modèle de langage par rapport au corpus de devtest. La section qui suit présente des résultats obtenus sur un corpus de parole broadcast news bien plus conséquent.

6.4 Expériences sur un corpus de parole broadcast news

N'ayant pas directement accès au corpus de parole broadcast news, Ebru Arisoy, de l'équipe de traitement de la parole de l'université de Bogaziçi, a eu la gentillesse de tester nos modèles de décomposition lexicaux. La méthodologie employée est celle qui est décrite dans l'article [Arisoy *et al.*, 2007]. Comme dans les expériences que nous avons menées sur l'amharique et le corpus de langue turque de parole lue, les morphes sont utilisés à la fois lors de l'apprentissage des modèles acoustiques de phones et lors de la création des modèles de langage.

Le corpus d'apprentissage audio comporte 71 heures d'émissions de radio et télévision transcrites manuellement, dont 68,6 heures sont utilisées pour l'entraînement des modèles acoustiques et 2,5 heures servent de corpus de devtest. Environ 7500 contextes sont modélisés par des triphones inter-mots à états liés (11 gaussiennes par état). Remarquons que ces modèles n'ont pas été ré-entraînés pour chaque modèle de décomposition alors qu'ils dépendent de la position inter-mots. Le fait d'utiliser le même jeu de modèles acoustiques pour toutes les représentations lexicales n'est pas optimal.

Le lexique de prononciations est fondé sur les graphèmes, aucune conversion graphème-phonème n'a été utilisée. Le lexique ne présente pas variante de prononciation, pour aucune des entrées lexicales.

Les modèles de langage sont des modèles *5-grammes* estimés sur des textes de sources variées totalisant 96,4M de mots, sauf pour le modèle fondé sur les mots qui est un modèle *3-grammes*. Les textes proviennent de livres, de journaux et magazines en ligne sur le

Web. Les modèles utilisés dans le décodage sont issus de l'interpolation d'un modèle de langage estimé sur les textes issus du Web et d'un modèle estimé sur les transcriptions qui totalisent 485k mots. Le coefficient d'interpolation du modèle fondé sur les transcriptions vaut 0,4 et est identique pour tous les systèmes testés, quelles que soient les options de décomposition. Aucun élagage (« pruning ») n'a été pratiqué sur les modèles de langage. Nous leur avons fourni les modèles de décomposition des mots en morphes, les mêmes que ceux que nous avons utilisés sur le corpus de parole lue de LDC. Ils ont réalisé les décompositions sur les lexiques issus du corpus de textes de 96,4M de mots et du corpus de transcriptions de 485k mots, avec l'algorithme Viterbi fourni dans le programme Morfessor.

Le décodeur utilisé est un décodeur en une seule passe, fourni par les laboratoires AT&T [Allauzen *et al.*, 2003]. Les treillis sont ensuite décodés par le modèle de langage pentagramme fondé sur les morphes. Les mots sont ensuite reconstitués avant de mesurer les performances des systèmes.

Le tableau 6.11 présente les taux d'erreurs de mots des différents systèmes. Tous les jeux d'options n'ont pas été testés, nous avons choisi les jeux au fur et à mesure que les résultats nous étaient donnés. Le meilleur système est le système MHTDC qui présente un WER de 37,6% soit un gain relatif de 5% par rapport au Baseline. Les performances des différents systèmes montrent que plus le nombre de mots décomposés est élevé, plus le WER diminue. Une corrélation entre le ratio du nombre de morphes divisé par le nombre de mots dans les textes d'apprentissage des modèles de langage et le WER semble établie. Le tableau 6.12 rappelle les WER et donne les valeurs de ce ratio pour les différents systèmes testés. Comme une balise WB est ajoutée pour représenter les frontières de mots dans les textes d'apprentissage, ces ratios sont tous supérieurs à deux. Le système MHTDC qui présente le WER le plus bas, a le ratio le plus grand, d'une valeur de 2,24.

Pour obtenir des ratios plus élevés, deux nouveaux modèles de décomposition ont été créés. Pour obtenir un nombre plus grand de décompositions, l'algorithme de Viterbi a été appliqué aux lexiques sans restriction de nombre d'occurrences minimal. Le lexique du corpus Bogaziçi comporte 517k mots distincts au lieu de 197k mots lorsqu'une limite inférieure de trois occurrences est appliquée. Ces deux modèles ont été construits pour les options MHTDC et MHTDCc, qui sont les options qui fournissent le plus grand nombre de décompositions. Les WER obtenus valent respectivement 37,4% et 38,1% avec des ratios valant 2,31 et 2,11.

La figure 6.1 illustre les chiffres du tableau 6.12, en donnant l'évolution du WER en fonction du ratio du nombre de morphes sur le nombre de mots. Le meilleur résultat présenté dans l'article [Arisoy *et al.*, 2007] a été intégré dans la courbe. Il s'agit d'un WER de 35,4%, obtenu avec un ratio de 2,38. Pour la génération des morphes de ce système, un lexique différent de celui que nous avons utilisé, comprenant 1,4M mots distincts, a été utilisé.

Dans ces expériences, aucune valeur seuil du ratio du nombre de morphes sur le nombre de

<i>Options</i>	<i># Morphes</i>	<i>WER (%)</i>
BL	50k	39,6
M	115,5k	39,2
M Cc	152,7k	39,9
MH	115,1k	39,2
MH Cc	151,5k	39,8
MH TDV	-	-
MH TDV Cc	153,0k	39,9
MH TDC	49,5k	37,6
MH TDC Cc	103,7k	38,2
MH TDCV	-	-
MH TDCV Cc	-	-

TAB. 6.11 – Nombre de morphes des lexiques, taux d’erreurs de mots (*WER*) des différents systèmes. Tous les systèmes n’ont pas été testés. NB : tous les *WER* sont donnés après recombinaison des sous-unités pour former des mots entiers.

<i>Options</i>	<i>Ratio</i>	<i>WER (%)</i>
BL	2,0	39,6
M	2,04	39,2
M Cc	2,02	39,9
MH	2,04	39,2
MH Cc	2,02	39,8
MH TDV	-	-
MH TDV Cc	2,02	39,9
MH TDC	2,24	37,6
MH TDC Cc	2,11	38,2
MH TDCV	-	-
MH TDCV Cc	-	-

TAB. 6.12 – Ratios du nombre de morphes divisé par le nombre de mots dans les textes d’apprentissage des modèles de langage des différents systèmes ($\#Morphes(incluant\ WB) / \#Mots$) et taux d’erreurs de mots (*WER*). Les systèmes qui n’ont pas été testés sont indiqués par un « - ». NB : tous les *WER* sont donnés après recombinaison des sous-unités pour former des mots entiers.

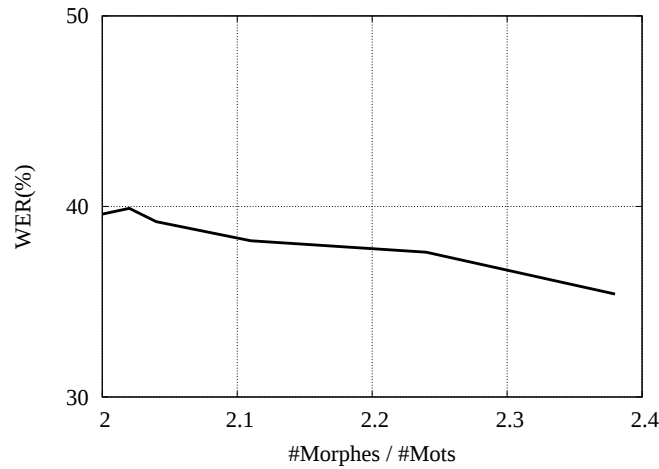


FIG. 6.1 – *Taux d'erreurs de mots en fonction du ratio du nombre de morphes divisé par le nombre de mots dans les textes d'apprentissage des modèles de langage des différents systèmes.*

mots n'a été mise en évidence. En effet, dans les expériences précédentes en amharique et en turc avec le corpus de parole lue, nous avons obtenu une dégradation des performances avec les modèles lexicaux qui décomposaient le plus grand nombre de mots. Plusieurs hypothèses pourraient expliquer cette différence.

Nous n'avons peut-être pas généré de systèmes qui décomposent suffisamment de mots pour observer ce seuil de confusion accrue. Le système de référence, fondé sur les mots, utilise un lexique de 50k mots, comprenant les mots apparaissant au moins cinq fois dans les textes. Comme dans toutes les expériences que nous avons menées, l'application des décompositions n'est pas réalisée sur ce petit lexique, mais sur un lexique contenant tous les mots, sans restriction de nombre d'occurrences. Comme le taux OOV initial avec le lexique de 50k mots est très élevé (il atteint 9,3%), le gain très important en couverture du fait des décompositions pourrait expliquer que nous n'observons pas de dégradation de performances.

Une autre explication pourrait être la différence de gestion des frontières de mots et la manière de reconstituer les mots entiers à partir des morphes. Dans les expériences précédentes, nous ajoutons un signe '+' en fin d'affixe pour pouvoir reconstituer les mots. L'équipe Busim préfère insérer une balise WB pour « Word Boundary » entre les séquences de morphes, pour signifier les frontières de mot. Cette méthode a l'avantage de ne pas différencier les formes préfixes des mots entiers pour les mots qui sont à la fois mot et morphe. Il reste à comparer les deux techniques pour tester cette hypothèse.

6.5 Conclusion

Dans ce chapitre, nous avons présenté des expériences de sélection automatique d'unités lexicales sur deux corpus différents en langue turque. Le premier corpus est un corpus de parole lue, de taille très réduite. Le second corpus est un corpus de parole de type « broadcast news », comprenant 70 heures de parole transcrite. Nous n'avons pas eu directement accès à ce corpus, l'équipe de traitement de la parole de l'université Bogaziçi d'Istanbul a mesuré les performances de leur système de transcription en utilisant les modèles lexicaux de décomposition que nous leur avons fournis.

Les systèmes de référence sont les systèmes fondés sur les mots entiers, auxquels sont comparées les performances des systèmes fondés sur les morphes. Sur le corpus de parole lue, le taux d'erreurs de mots (WER) du système de référence atteint 35,9%. Sur le corpus de parole broadcast news, le WER est de 39,6%. De légers gains ont été obtenus avec certains systèmes fondés sur les morphes, les gains les plus élevés étant de 5% relatifs pour les expériences sur les deux types de corpus.

Les tendances observées dans les expériences menées sur les deux corpus sont différentes. Avec le corpus de parole Broadcast news, plus le nombre de mots décomposés est grand, meilleures sont les performances obtenues. Une certaine corrélation semble exister entre le WER et le ratio du nombre de morphes divisé par le nombre de mots dans les textes qui servent à estimer les modèles de langage. Plus ce ratio est grand (dans nos expériences, il varie entre 2,0 et 2,31), moins les WER des systèmes sont élevés. En revanche, les expériences menées sur le corpus de parole lue montrent globalement les mêmes tendances que les expériences menées sur l'amharique, à savoir l'intérêt de limiter les décompositions avec une contrainte de génération des morphes. Sur ce corpus, le meilleur système est celui qui utilise l'algorithme Morfessor de départ avec la contrainte fondée sur les alignements phonotactiques. La propriété des traits distinctifs des consonnes a permis de réduire d'un facteur trois la taille des lexiques, néanmoins les performances obtenues avec cette propriété sont moins bonnes et surtout une légère dégradation est observée avec l'emploi de la contrainte Cc. L'interprétation des résultats sur ce corpus est limitée d'une part par la taille très réduite de ce corpus, et d'autre part par l'inadéquation manifeste entre les textes issus de textes d'information collectés sur le Web servant à l'apprentissage des modèles de langage, et le corpus de parole lue, constitué de phrases phonétiquement équilibrées.

Les gains de performance observés avec l'augmentation du ratio entre le nombre de morphes et le nombre de mots peuvent être dus aux caractéristiques de la langue turque elle-même, à sa morphologie bien particulière, mais également à la méthodologie légèrement différente des expériences que nous avons menées sur l'amharique et sur le corpus de parole lue en langue turque. Notamment l'usage d'une balise marquant les frontières de mots permet de ne pas différencier les formes morphes des mots qui sont à la fois préfixe et mot. Enfin nous avons remarqué que les taux d'erreurs de graphèmes des systèmes fondés sur les morphes sont tous meilleurs que celui du système fondé sur les mots, ce

qui indique que beaucoup d'erreurs sont liées à la recombinaison des morphes entre eux pour reformer des mots entiers.

Conclusion et perspectives

Les technologies liées à la parole, synthèse et reconnaissance, sont de plus en plus présentes dans les pays dits développés. Au cours de cette thèse, nous nous sommes intéressés aux langues qui ne possèdent pas de grandes quantités de ressources numériques, textes et audio, nécessaires au développement de technologies vocales et plus spécifiquement de systèmes de reconnaissance de la parole. La réduction du temps et des moyens nécessaires à l'adaptation des systèmes à de nouvelles langues est devenue une question centrale. Les travaux réalisés au cours de cette thèse s'inscrivent dans cette démarche, avec la volonté d'améliorer les performances de systèmes construits pour des langues peu dotées, en utilisant les techniques qui sont celles des systèmes à l'état de l'art.

Dans un premier temps, nous avons précisé la définition des langues peu dotées. Pour ces langues, les ordres de grandeur caractéristiques des corpus de données disponibles ou accessibles plus ou moins rapidement sont de l'ordre du million de mots pour les textes, et de quelques heures de parole transcrites manuellement. En outre, nous avons caractérisé les langues peu dotées plus par rapport aux problèmes liés au manque de textes, dû à une production écrite numérisée faible. Nous avons mis en avant le problème du manque de textes en nous appuyant sur quelques études de la littérature sur l'influence de la quantité de données audio et/ou textes sur les performances des systèmes de reconnaissance. Ces études ont montré que des transcriptions fines manuelles ne sont pas obligatoires pour obtenir des modèles acoustiques performants. Des techniques itératives d'apprentissage des modèles acoustiques à partir de transcriptions automatiques, peuvent mener à des performances comparables à celles obtenues avec des transcriptions manuelles.

Pour illustrer les problèmes rencontrés lors du développement d'un système de reconnaissance pour une langue peu dotée, nous avons travaillé sur deux langues : l'amharique, langue officielle de l'Éthiopie, et le turc. Jusqu'à très récemment, les corpus de parole en turc étaient très limités. Néanmoins, les recherches très actives et le dynamisme de cette langue sur le Web, feront que d'ici quelques années le turc sera une langue bien dotée. L'amharique, en revanche, souffre d'une présence très restreinte sur Internet, qui n'est peut-être pas favorisée par le syllabaire spécifique qu'elle utilise, pour lequel les outils informatiques habituels n'existent pas, ou restent très limités.

Outre les problèmes de normalisation des textes (translittération, transformation des

chiffres et dates en mots, formes orthographiques variées pour un seul mot, etc. . .), l'un des premiers problèmes rencontrés est l'élaboration d'un lexique de prononciations. La méthode la plus simple est l'approche graphémique, qui consiste à associer un phonème à chaque graphème. Pour l'amharique, la relation entre l'écrit et l'oral est relativement simple, et il nous a été possible d'associer à chaque symbole une lettre ou une suite de lettres latines qui représente les phonèmes correspondants. Dans le but d'évaluer une méthode automatique d'identification de variantes de prononciations, un lexique constitué des phonèmes de la langue, autorisant la substitution de chaque phonème par n'importe quel autre phonème de la langue, a été construit. Ce lexique a été utilisé pour aligner les transcriptions de référence aux données audio au niveau phonotactique. Les substitutions de phonèmes les plus fréquentes sont sélectionnées et utilisées pour obtenir quelques variantes pour chaque mot. La pertinence des variantes de prononciation peut être estimée en mesurant les taux d'alignement des mots avec chacune des variantes potentielles. L'utilisation de variantes pertinentes améliore la qualité des modèles acoustiques, comme en témoignent les augmentations de vraisemblance, et la diminution du nombre de segments rejetés dans la phase d'entraînement des modèles acoustiques.

Nous avons ensuite cherché à caractériser l'influence de la quantité de textes et de données d'apprentissage audio sur les performances d'un système de reconnaissance d'architecture standard. Le taux d'erreurs de mots (WER) du système entraîné sur la totalité des corpus audio (35h) atteint 24,4% pour un taux de mots hors-vocabulaire (OOV) très élevé de 7,0% (entre 3 à 10 fois supérieur à un taux usuel pour une langue bien dotée), avec un lexique comprenant 133k mots. En réduisant à 10h de données audio d'apprentissage pour les modèles acoustiques, le WER atteint 26,7%, soit une augmentation de seulement 9,4% relatifs. Avec seulement 2h de données d'apprentissage audio, le taux d'erreurs double. Des expériences complémentaires sur l'influence des modèles de langage comparée à l'influence des modèles acoustiques ont été menées. Les tendances discernées dans ces études semblent indiquer qu'avec très peu de données audio, typiquement moins de 10h, l'influence de la quantité des textes est moindre comparée à l'influence des modèles acoustiques. Cette tendance semble s'inverser pour des corpus audio plus grands. Il apparaît que les différences de performances entre des systèmes entraînés sur 10h et sur 35h sont proches, lorsqu'un corpus de textes d'au moins un million de mots est utilisé pour l'entraînement des modèles de langage. Ce résultat suggère qu'à partir de ce point de fonctionnement (10h de données audio transcrites et 1M de mots), collecter de nouvelles quantités de textes pourrait être plus efficace que transcrire quelques heures de données audio supplémentaires.

Le manque de textes est l'un des principaux facteurs à l'origine de taux d'OOV élevés pour les langues peu dotées. La réduction de ces taux, sources de plus d'une erreur de reconnaissance par mot hors-vocabulaire en moyenne, nous a paru être le premier problème général à étudier pour les langues peu dotées. Nous avons donc axé nos recherches sur la modélisation lexicale, en cherchant à développer des outils indépendants de la langue, pour sélectionner des unités lexicales optimales pour la reconnaissance de

la parole grand vocabulaire. Ces unités peuvent être les mots eux-mêmes, ou bien des sous-unités, pour lesquelles nous avons repris le nom de « morphes », couramment utilisé dans la littérature. Une première expérience encourageante de décompositions à l'aide d'un petit jeu d'affixes choisis arbitrairement parmi les plus fréquents avait montré un léger gain de performances de 3,1% relatifs pour la langue amharique.

Nous avons utilisé des méthodes non-supervisées dans la mesure où elles s'inscrivent dans des paradigmes statistiques qui limitent la dépendance à la langue étudiée. D'autre part, les méthodes non-supervisées ne nécessitent pas ou peu de connaissances linguistiques expertes, et elles sont facilement modifiables. Une fois que les sous-unités sont identifiées, plusieurs possibilités d'utilisation au sein d'un système sont possibles. Elles peuvent être utilisées dans toutes les étapes du décodeur —les modèles de langage et les modèles acoustiques sont alors entraînés sur des textes et des transcriptions dont les mots ont été décomposés— ou alors uniquement au niveau post-traitement, c'est-à-dire à la dernière étape de décodage qui est l'étape de rescoring ou de réestimation des scores des treillis de mots, pour obtenir la meilleure hypothèse. Dans ce dernier cas, un modèle de langage fondé sur les sous-unités est utilisé pour rescorer les hypothèses issues d'un système fondé sur les mots entiers. La littérature ne permet pas de savoir, de manière générale, quelle est la meilleure méthode, quels que soient la langue et le type de parole. Néanmoins les études les plus récentes utilisent la première méthode, avec des résultats qui semblent globalement cohérents, et pour cette raison nous avons opté pour la première méthode.

Notre choix de départ s'est porté sur l'algorithme de type maximisation de vraisemblance appelé « Morfessor », développé à l'université finlandaise d'Helsinki. Il présente l'avantage, par rapport à d'autres algorithmes disponibles, de ne faire aucune hypothèse sur le nombre de sous-unités qui peuvent composer les mots. De ce fait, cet algorithme ne dépend pas de la langue étudiée. Les propriétés lexicales qu'il prend en compte sont uniquement fondées sur les formes écrites des mots, comme les suites de caractères, les nombres d'occurrences. Nous avons proposé de nouvelles propriétés, pour tenter d'introduire des propriétés d'ordre oral dans les modèles de décompositions des mots. Les modifications du programme de départ rendent l'algorithme plus dépendant de la langue à cause de l'utilisation du jeu de phones et de traits distinctifs spécifiques, néanmoins il est très simple de les adapter à une nouvelle langue. L'une de ces propriétés, basée sur les traits distinctifs des consonnes et des voyelles de la forme phonémique principale associée à chaque mot, tente d'apporter une mesure de dissimilarité acoustico-phonétique entre les sous-unités trouvées. Cette propriété est d'ordre théorique, dans la mesure où les traits distinctifs sont des traits binaires issus de la littérature de phonétique générale. Nous avons introduit également une contrainte empirique à partir d'un corpus de parole aligné, qui essaie de limiter l'accroissement de confusion acoustico-phonétique dû à l'introduction d'unités de reconnaissance de petite taille. À partir d'alignements phonotactiques avec un lexique constitué des seuls phones de la langue, sont déterminées les paires de phones qui se substituent le plus l'un à l'autre. Les décompositions qui génèrent des sous-unités qui ne diffèrent que de deux phones qui forment une paire identifiée

précédemment comme une paire de substitution, sont interdites. D'autres modifications ont été apportées à l'algorithme de départ, comme par exemple l'introduction d'un algorithme inspiré de l'algorithme de Harris de découverte de frontière de morphes. Il permet d'identifier plus de décompositions que ne le fait l'algorithme Morfessor de départ.

De manière générale pour les tests menés sur l'amharique, de petits gains compris entre 0,5% et 2,0% relatifs ont été obtenus pour les systèmes utilisant entre autres la contrainte visant à limiter la confusion entre les unités. Sans cette contrainte, les lexiques obtenus sont de taille plus petite (la taille est même divisée par deux pour certains modèles), mais la confusion acoustico-phonétique entre petites unités semble plus importante, et les systèmes correspondants présentent des WER plus grands que le WER de référence. Le système qui a le WER le plus petit utilise la contrainte de sélection des morphes, mais également la propriété liée aux traits distinctifs des voyelles. Néanmoins, les différences des scores des systèmes sont trop petites pour pouvoir affirmer que la propriété des traits distinctifs est réellement nécessaire.

Des expériences sur deux corpus distincts de parole en turc ont montré des résultats sensiblement différents. Sur le corpus de parole lue fourni par LDC, comprenant un peu moins de 5 heures d'audio, un gain de 5% relatifs a été obtenu, avec le système basé sur les morphes identifiés par l'algorithme Morfessor de départ, avec uniquement l'ajout de la contrainte Cc, fondée sur les alignements phonotactiques du corpus d'apprentissage des modèles acoustiques. La propriété des traits distinctifs a montré des performances légèrement moins bonnes, et l'application de la contrainte Cc a fait contre-emploi sur les systèmes utilisant les traits distinctifs sur les consonnes, qui sont les systèmes qui présentent le plus grand nombre de décompositions, avec des lexiques de taille trois fois plus petite que le lexique de départ. Nous avons également mesuré les taux d'erreurs de lettres (LER), et il est apparu que les LER des systèmes fondés sur les morphes sont tous moins élevés que le LER du système de référence fondé sur les mots. Cela signifie qu'un grand nombre d'erreurs proviennent des recombinaisons des morphes en mots entiers. Les résultats obtenus sur ce corpus sont cependant peu fiables, en raison de l'inadéquation manifeste des modèles de langage vis-à-vis du corpus de devtest. Les modèles de langage ont été estimés sur des corpus de textes d'information, alors que le corpus devtest est composé de phrases artificielles, créées pour être phonétiquement équilibrées.

Les expériences menées sur un second corpus présentent des résultats encourageants. Il s'agit d'un corpus de 70 heures de parole provenant d'émissions d'information de radios et télévisions (broadcast news). Ce corpus est la propriété de l'équipe de traitement de la parole « BUSIM », de l'université stambouliote de Bogazçi, qui a réalisé les expériences avec les modèles lexicaux que nous leur avons donnés. Les modèles de décompositions qui ont été testés sur ce corpus sont les mêmes que ceux qui ont été utilisés sur le corpus de parole lue. Le résultat intéressant qui est apparu avec ce corpus, est la corrélation qui a été mise en évidence entre le ratio du nombre de morphes divisé par le nombre de mots des textes. Plus ce ratio est grand, c'est-à-dire plus le nombre de mots décomposés est grand, plus les performances des systèmes de reconnaissance sont meilleures. Le seuil du nombre

de décompositions maximal, qui conduirait à un accroissement du taux d'erreurs, n'a pas été atteint avec les modèles que nous avons générés. Lors des expériences précédentes en amharique et sur le corpus de parole lue en turc, nous avons mis en évidence le problème de la recombinaison des morphes entre eux pour former des mots entiers, qui semble être une source de nombreuses erreurs. L'utilisation d'un signe '+' accolé aux affixes peut entraîner des confusions pour les mots du lexique qui sont à la fois affixe et à la fois mot. Le système de l'équipe Busim procède différemment : une balise WB représente les frontières de mots, et le lexique ne fait pas de distinction entre les morphes et les mots. Cette méthode permettrait peut-être de limiter les confusions lors des recombinaisons des morphes. La balise de frontière de mot, qui fait partie des mots appelés « non-événement », c'est-à-dire des mots qui n'ont pas de prononciation associée, comme la balise < s > par exemple, semble efficace. Cependant, elle nécessite d'être capable de définir une prononciation nulle, ce qui peut poser problème pour le décodage.

La sélection automatique d'unités lexicales adaptées à la reconnaissance de la parole nous a semblé être un point très important dans l'élaboration d'un système pour une langue peu dotée, pour diminuer le nombre de mots hors-vocabulaire, et éventuellement améliorer la représentation des mots peu fréquents lors de l'estimation des modèles de langage. Si des gains ont été obtenus sur des corpus de langues différentes et de types de parole différents, de nombreux points restent à explorer. Par exemple, les méthodes de recombinaison des sous-unités entre elles pour reformer des mots entiers restent à comparer.

D'autre part, nous avons travaillé sur le premier problème qui se pose lors de l'élaboration d'un nouveau système pour une langue peu dotée, à savoir la diminution du taux de mots inconnus. Le problème suivant, qui se cache derrière les forts taux de mots hors-vocabulaire, est la faiblesse du pouvoir de prédictibilité des modèles de langage. La modélisation lexicale, avec les modèles statistiques de décompositions des mots en unités plus petites en particulier, peut être vue comme une première étape d'amélioration des systèmes de reconnaissance pour les langues peu dotées, mais elle ne résout pas le problème de fond du manque de textes, qui reste donc à aborder du point de vue de la modélisation syntaxique.

Les recherches menées au cours de cette thèse ont concerné principalement la modélisation lexicale et la création des lexiques de prononciations. En ce qui concerne les modèles acoustiques, il faut souligner les perspectives très prometteuses de l'entraînement non-supervisé des modèles. Quelques études ont montré le potentiel de méthodes qui permettent de se passer de transcriptions manuelles fines. Ce type d'apprentissage serait très intéressant à développer pour les langues peu dotées.

Enfin nous avons étudié deux langues peu dotées, qui disposent d'un minimum de données, et qui sont amenées à devenir de manière certaine des langues bien dotées dans un avenir proche. Qu'en est-il des langues très peu ou pas dotées, comme les langues de tradition orale ? Les dialectes arabes par exemple, qui ne possèdent pas de système de transcription écrite, seront sûrement l'objet de travaux de recherche dans les prochaines années. Sans

nul doute le panel de langues différentes faisant l'objet de recherches en traitement de la parole, va s'élargir très vite, le besoin de nouvelles technologies se faisant grandissant dans des pays de plus en plus nombreux.

Annexe A

Le tableau 6.13 donne la table complète des correspondances entre les symboles geez et les caractères latins choisis en interne au LIMSI-CNRS.

La lettre 'ʔ' a été choisie pour représenter le coup de glotte. Tous les symboles de type CV (consonne voyelle) sont donnés dans la première colonne, les symboles de type CwV où w est une semi-consonne sont donnés dans la seconde colonne. Seules les syllabes avec le noyau vocalique /E/ sont données mais chaque symbole est légèrement modifié pour indiquer les 6 autres voyelles (les voyelles sont appelées « ordres »). Presque tous les symboles de type CV ont un homologue CwV mais pas pour tous les ordres. Les symboles indiqués avec la voyelle /a/ n'existent pas aux autres ordres (seule la forme avec la voyelle /a/ existe). Ceux indiqués avec la voyelle /E/ existent aux autres ordres.

TAB. 6.13 – Syllabaire geez donné avec les transcriptions en caractères latins. La première colonne donne les formes CV du premier ordre (voyelle translittérée par 'E') et la seconde donne les formes de type Cwv où w est une semi-consonne. Tous les symboles geez listés ici possèdent six autres formes pour les six autres voyelles sauf les formes de type CwV données avec la voyelle 'a' qui n'existent qu'avec cette voyelle.

ሀ	hE	ሐ	hwa
ለ	lE	ላ	lwa
ሐ	hE	ሐ	hwa
ሞ	mE	ሞ	mwa
ሥ	sE	ሥ	swa
ር	rE	ረ	rwa
ሰ	sE	ሰ	swa
ሸ	SE	ሸ	Swa
ቅ	KE	ቅ	KwE
ቆ	KE	ቆ	KwE
ብ	bE	ብ	bwE
ቨ	bE	ቨ	vwa
ተ	tE	ተ	twa
ቸ	CE	ቸ	Cwa
ኀ	hE	ኀ	hwE
ኃ	nE	ኃ	nwa
ኘ	NE	ኘ	Nwa
ኀ	?E	ኀ	?wE
ከ	kE	ከ	kwE
ከ	hE	ከ	hwE
ወ	wE	-	-
ኀ	?E	-	-
ዝ	zE	ዝ	zwa
ዘ	ZE	ዘ	Zwa
ይ	jE	ይ	jwE
ደ	dE	ደ	dwa
ደ	dE	ደ	dwa
ጅ	JE	ጅ	Jwa
ገ	gE	ገ	gwE
ጥ	TE	ጥ	Twa
ቸ	QE	ቸ	Qwa
ቶ	PE	ቶ	Pwa
ቶ	tsE	ቶ	tswa
ፅ	tsE	-	-
ፍ	fE	ፍ	fwE
ፐ	pE	ፐ	pwE

Annexe B

Les tableaux 6.14, 6.15 et 6.16 donnent respectivement les tables des traits distinctifs des voyelles et des consonnes de la langue amharique tels qu'ils ont été utilisés pour estimer les propriétés liées aux traits distinctifs dans l'algorithme de Morfessor modifié.

De la même manière, les tableaux 6.17, 6.18 et 6.19 donnent respectivement les tables des traits distinctifs des voyelles et des consonnes de la langue turque.

Tab. 6.17 – Sous-ensemble de traits distinctifs utilisés pour les voyelles du turc

Trait	voyelles							
Graphème	a	e	i	u	ü	ɪ	o	ö
Phone LIMSI	a	e	i	u	y	x	o	@
symbole API	a	e	i	u	y	ɪ	o	ø
haut	0	0	1	1	1	0	0	0
bas	1	1	0	0	0	0	1	1
arrondi	0	0	0	1	1	0	1	1
réduit	0	0	0	0	0	1	0	0
antérieur	0	1	1	0	1	0	0	0

TAB. 6.18 – Traits distinctifs des consonnes turques. Lieu : L labial, D dental, A alvéolaire, P palatal, V vélaire, U uvulaire

Trait	consonnes									
Phone set LIMSI	b	d	g	p	t	k	C	J	v	
Graphèmes	b	d	g	p	t	k	ç	c	v	
API	b	d	g	p	t	k	ʃ	dʒ		
voisé/non-voisé	1	1	1	0	0	0	0	1	1	1
lieu	L	D	V	L	D	V	P	P	V	
sonore	0	0	0	0	0	0	0	0	0	0
glottalisé	0	0	0	0	0	0	0	0	0	0
coronal	0	1	0	0	1	0	1	1	0	0
antérieur	1	1	0	1	1	0	0	1	0	0
distribué	1	0	0	1	0	0	0	0	0	0
haut	0	0	1	0	0	1	0	0	0	0
arrière	0	0	1	0	0	1	0	0	0	0
arrondi	0	0	0	0	0	0	0	0	0	0
continu	0	0	0	0	0	0	0	0	1	1
latéral	0	0	0	0	0	0	1	0	0	0
nasal/oral	0	0	0	0	0	0	0	0	0	0
strident	0	0	0	0	0	0	1	1	1	1
affriqué	0	0	0	0	0	0	1	1	0	0

TAB. 6.19 – Traits distinctifs des consonnes turques (suite). Lieu : L labial, D dental, A alvéolaire, P palatal, V vélaire, U uvulaire

[illegible]

Bibliographie

- [Abate & Menzel, 2007] S.T. Abate & W. Menzel. Automatic Speech Recognition for an Under-Resourced Language - Amharic. In *Proceedings of Interspeech*, pages 1541–1544, Antwerp, 2007.
- [Abate *et al.*, 2005] S.T. Abate, W. Menzel, & B. Tafila. An amharic speech corpus for large vocabulary continuous speech recognition. In *Proceedings of Interspeech*, Lisboa, 2005.
- [Abdou, 2004] S. et al Abdou. The 2004 BBN Levantine Arabic and Mandarin CTS Transcription Systems. In *DARPA RT-04 Workshop*, New York, 2004.
- [Adda *et al.*, 1997] G. Adda, M. Adda-Decker, J.L. Gauvain, & L. Lamel. Text normalization and speech recognition in French. In *Proceedings of EuroSpeech*, volume 5, pages 2711–2714, Rhodes, 1997.
- [Adda-Decker & Adda, 2000] M. Adda-Decker & G. Adda. Morphological decomposition for ASR in German. In *Proceedings Workshop on Phonetics and Phonology in ASR*, volume PHONUS 5, pages 129–143, Saarbrücken, 2000.
- [Adda-Decker & Lamel, 1999] M. Adda-Decker & L. Lamel. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29 :83–98, 1999.
- [Adda-Decker, 2003] M. Adda-Decker. A corpus-based compounding algorithm for German lexical modeling in LVCSR. In *Proceedings of Eurospeech*, Geneva, 2003.
- [Allauzen *et al.*, 2003] C. Allauzen, M. Mohri, & M.D. Riley. DCD Library - Decoder Library <http://www.research.att.com/sw/tools/dcd>. ATT Labs - Research, 2003.
- [Amorrortu *et al.*, 2004] E. Amorrortu, A. Barreña, I. Idiazabal, E. Izagirre, P. Ortega, & B. Uranga. *World Languages Review Synthesis*. Unesco Etxea, 2004.
- [Appleyard, 1995] D. Appleyard. *Colloquial Amharic*. Routledge, London, 1995.
- [Arisoy & Saraclar, 2006] E. Arisoy & M. Saraclar. Lattice Extension and Rescoring Based Approaches for LVCSR of Turkish. In *Proceedings of ICSLP*, Pittsburg, 2006.
- [Arisoy *et al.*, 2007] E. Arisoy, H. Sak, & M. Saraclar. Language Modeling for Automatic Turkish Broadcast News Transcription. In *Proceedings of Interspeech*, Antwerp, 2007.

- [Asker *et al.*, 2007] L. Asker, A. Argaw, B. Gambäck, & M. Sahlgren. Applying Machine Learning to Amharic Text Classification. In *Proceedings of the 5th World Congress of African Linguistics*, Köln, Rüdiger Köppe Verlag, 2007.
- [Bahl *et al.*, 1976] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, & H.F. Silverman. Preliminary results on the performance of a system for the automatic recognition of continuous speech. In *Proceedings of ICASSP*, Philadelphia, 1976.
- [Bahl *et al.*, 1991] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, & M. A. Picheny. Context-dependent modeling of phones in continuous speech using decision trees. In *DARPA Workshop on Speech and Natural Language*, pages 264–269, 1991.
- [Baum *et al.*, 1970] L. E. Baum, T. Petrie, G. Soules, & N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41 :164–171, 1970.
- [Beesley & Karttunen, 2003] K.R. Beesley & L. Karttunen. *Finite State Morphology*. CSLI Publications, 2003.
- [Boula de Mareüil, 1997] P. Boula de Mareüil. *Étude linguistique appliquée à la synthèse de la parole à partir du texte*. PhD thesis, Université Paris XI Orsay, 1997.
- [Chou & Juang, 2003] W. Chou & F. Juang. *Pattern Recognition in Speech and Language Processing*. CRC Press, 2003.
- [Choueiter *et al.*, 2007] G. Choueiter, S. Seneff, & J. Glass. New Word Acquisition Using SubWord Modeling. In *Proceedings of Interspeech*, pages 1765–1768, Antwerp, 2007.
- [Clews, 1997] J. Clews. Digital Language Access : Scripts, Transliteration, and Computer Access. In *The Magazine of Digital Library Research*, International Organization for Standardization, Conversion of written languages ISO/TC46/SC2, 1997.
- [Creutz & Lagus, 2005] M. Creutz & K. Lagus. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0. *Computer and Information Science*, Report A81 :27, 2005.
- [Creutz *et al.*, 2007] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pytkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, & A. Stolcke. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Proceedings of NAACL-HLT*, pages 380–387, Rochester, 2007.
- [Crystal, 2000] D. Crystal. *Language Death*. Cambridge University Press, 2000.
- [Davis & Mermelstein, 1980] S. Davis & P. Mermelstein. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoustics, Speech & Signal Processing*, 28(4) :357–366, 1980.
- [Demisse & Imbert-Vier, 1996] D. Demisse & S. Imbert-Vier. *Amharique pour franco-phones*. L'Harmattan, Paris, 1996.

- [Denoual & Lepage, 2006] E. Denoual & Y. Lepage. The character as an appropriate unit of processing for non-segmenting languages. In *Proceedings of the 12th Annual Meeting of The Association of NLP*, pages 731–734, 2006.
- [Emerson, 2005] T. Emerson. The second international chinese word segmentation ba-keoff. In *4th SIGHAN Workshop on Chinese Language Processing*, Jeju, 2005.
- [Engelbrecht & Schultz, 2005] H. Engelbrecht & T. Schultz. Rapid Development of an Afrikaans-English Speech-to-Speech Translator. In *Proceedings of International Workshop of Spoken Language Translation (IWSLT)*, Pittsburgh, 2005.
- [Erdogan *et al.*, 2005] H. Erdogan, O. Buyuk, & K. Oflazer. Incorporating language constraints in sub-word based speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun, 2005.
- [Eyassu & Gamback, 2005] S. Eyassu & B. Gamback. Classifying Amharic news texts using self-organizing Maps. In *ACL05 Workshop on computational Approaches to Semitic Languages*, Ann Arbor, 2005.
- [Fissaha & Haller, 2003] S. Fissaha & J. Haller. Amharic verb lexicon in the context of machine translation. In *Proceedings of TALN*, Batz-sur-Mer, 2003.
- [Gauvain *et al.*, 2002] J.L. Gauvain, L. Lamel, & G. Adda. The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2) :89–108, 2002.
- [Geutner *et al.*, 1998] P. Geutner, M. Finke, & A. Waibel. Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multilingual broadcast news. In *Proceedings of ICSLP*, Sydney, 1998.
- [Geutner *et al.*, 2000] P. Geutner, C. Cariki, & T. Schultz. Towards better speech recognition for agglutinative languages. In *Proceedings of ICASSP*, Istanbul, 2000.
- [Geutner, 1995] P. Geutner. Using morphology towards better large-vocabulary speech recognition systems. In *Proceedings of ICASSP*, pages 445–448, 1995.
- [Grimes, 1996-2000] B.F. Grimes. Ethnologue : Languages of the World. In *Summer Institute of Linguistics*, Dallas, 1996-2000.
- [Hacioglu *et al.*, 2003] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, & M. Creutz. On lexicon creation for Turkish LVCSR. In *Proceedings of Eurospeech*, pages 1165–1168, Geneva, 2003.
- [Hagège, 2000] C. Hagège. *Halte à la mort des langues*. Odile Jacob, 2000.
- [Hagège, 2005] C. Hagège. *Words And Worlds : World Languages Review*. Multilingual Matters, 2005.
- [Halle & Clements, 1983] M. Halle & G.N. Clements. *Problem Book in Phonology*. The MIT Press, 1983.
- [Harris, 1955] Z.S. Harris. From phoneme to morpheme. *Language*, 31 :190–222, 1955.
- [Hermansky, 1990] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4) :1738–1752, 1990.

- [Isenberg, 1842] Rev. C.W. Isenberg. *Grammar of the Amharic Language*. AES Publications, London, 1842.
- [Iyer *et al.*, 1997] R. Iyer, M. Ostendorf, & M. Meteer. Analyzing and predicting language model improvements. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.
- [Jakobson *et al.*, 1952] R. Jakobson, G. Fant, & M. Halle. *Preliminaries to Speech Analysis*. MIT Press, Cambridge, 1952.
- [Jelinek, 1976] F. Jelinek. Continuous Speech Recognition by Statistical Methods. In *IEEE*, volume 64 :4, pages 532–556, 1976.
- [Kanthak & Ney, 2002] S. Kanthak & H. Ney. Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition. In *Proceedings of ICASSP*, volume 1, pages 845–848, Orlando, 2002.
- [Kershaw *et al.*, 1996] D. Kershaw, A.J. Robinson, & S.J. Renals. The 1995 Abbot hybrid connectionist-HMM large-vocabulary recognition system. *Proceedings of ARPA Speech Recognition Workshop*, pages 93–98, 1996.
- [Kiecza *et al.*, 1999] D. Kiecza, T. Schultz, & A. Waibel. Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. In *Proceedings of ICSP*, pages 323–327, Seoul, 1999.
- [Killer *et al.*, 2003] M. Killer, S. Stüker, & T. Schultz. Grapheme based speech recognition. In *Proceedings of Eurospeech*, Genf, 2003.
- [Kirchhoff & Sarikaya, 2007] K. Kirchhoff & R. Sarikaya. Processing morphologically rich languages. In *Workshop of Interspeech*, Antwerp, 2007.
- [Kirchhoff *et al.*, 2002] K. Kirchhoff, J. Bilmes, J. Henderson, & R. Schwartz. Novel speech recognition models for arabic. In *Johns-Hopkins University Summer Research Workshop*, volume Final report, 2002.
- [Kneser & Ney, 1995] R. Kneser & H. Ney. Improved backing-off for M-gram language modeling. In *Proceedings of ICASSP*, volume 1, pages 181–184, Detroit, 1995.
- [Kurimo *et al.*, 2006] M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, & M. Saraclar. Unsupervised segmentation of words into morphemes – morpho challenge 2005 : Application to automatic speech recognition. In *Proceedings of ICSLP*, Pittsburg, 2006.
- [Lamel & Adda, 2000] J-L. Lamel, L. and Gauvain & G. Adda. Lightly supervised acoustic model training. *ISCA ITRW Workshop on Automatic Speech Recognition : Challenges for the new Millenium*, pages 150–154, 2000.
- [Lamel & Adda, 2002] J-L. Lamel, L. and Gauvain & G. Adda. Unsupervised acoustic model training. In *Proceedings of ICASSP*, pages 877–880, Orlando, 2002.
- [Lavoie & O’Neill, 1998-2003] B.F. Lavoie & E.T. O’Neill. How ”World Wide” is the Web ? Technical report, Annual review of OCLC, 1998-2003.
- [Le *et al.*, 2006] V.-B. Le, L. Besacier, & T. Schultz. Acoustic-Phonetic Unit Similarities for Context-Dependent Acoustic Model Portability. In *Proceedings of ICASSP*, Toulouse, 2006.

- [Le, 2006] V.-B. Le. *Reconnaissance automatique de la parole pour des langues peu dotées*. PhD thesis, Université Joseph Fourier, Grenoble, 2006.
- [Lee *et al.*, 1990] H. Lee, L. R. Rabiner, R. Pieraccini, & J. G. Wilpon. Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4(2) :127–165, 1990.
- [Martin *et al.*, 1997] S.C. Martin, J. Liermann, & H. Ney. Adaptive topic dependent language modelling using wordbased varigrams. In *Proceedings of Eurospeech*, volume 3, pages 1447–1450, Rhodes, 1997.
- [Matthews, 1991] P. H. Matthews. *Morphology*. Cambridge University Press, Cambridge, 1991.
- [Messaoudi *et al.*, 2006] A. Messaoudi, J-L. Gauvain, & L. Lamel. Arabic broadcast news transcription using a one million word vocalized vocabulary. In *Proceedings of ICASSP*, volume I, pages 1093–1096, Toulouse, 2006.
- [Mircheva, 2006] A. Mircheva. Bulgarian speech recognition and multilingual language modeling. Technical report, Universität Karlsruhe, 2006.
- [Moore, 2003] R. K. Moore. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech*, pages 2582–2584, Geneva, 2003.
- [Nakajima *et al.*, 2000] H. Nakajima, Y. Sagisaka, & H. Yamamoto. Pronunciation variants description using recognition error modeling with phonetic derivation hypotheses. In *Proceedings of ICSLP*, volume 3, pages 1093–1096, Beijing, 2000.
- [Nguyen *et al.*, 2004] S. Nguyen, L. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.L. Gauvain, G. Adda, H. Schwenk, & F. Lefevre. The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System. In *Proceedings of DARPA RT04*, Palisades NY, 2004.
- [Nguyen *et al.*, 2005] L. Nguyen, B. Xiang, M. Afify, S. Abdou, S. Matsoukas, R. Schwartz, & J. Makhoul. The BBN RT04 English Broadcast News Transcription System. In *Proceedings of ICASSP*, pages 1673–1676, Lisbon, 2005.
- [Nimaan *et al.*, 2007] A. Nimaan, P. Nocera, F. Bechet, & J-F. Bonastre. Information retrieval strategies for accessing african audio corpora. In *Proceedings of Interspeech*, 2007.
- [Nimaan, 2007] A. Nimaan. *Sauvegarde du patrimoine oral africain : conception de système de transcription automatique de langues peu dotées pour l'indexation des archives audio*. PhD thesis, Université d'Avignon et des pays du Vaucluse, 2007.
- [O'Neill *et al.*, 1997] E.T. O'Neill, P.D. McClain, & B.F. Lavoie. A methodology for sampling the World Wide Web. Technical report, Technical Report, OCLC Annual Review of Research, 1997.
- [Pallett, 1990] D. et al. Pallett. Tools for the analysis of benchmark speech recognition tests. In *Proceedings of ICASSP*, volume 1, pages 97–100, 1990.

- [Park *et al.*, 2005] A. Park, T. Hazen, & J. Glass. Automatic processing of audio lectures for information retrieval : Vocabulary selection and language modeling. In *Proceedings of ICASSP*, Philadelphia, 2005.
- [Pellegrini & Lamel, 2006] T. Pellegrini & L. Lamel. Experimental detection of vowel pronunciation variants in Amharic. In *Proceedings of LREC*, Genoa, 2006.
- [Rabiner & Juang, 1986] L.R. Rabiner & B.H. Juang. An introduction to Hidden Markov Models. *IEEE Acoustics Speech and Signal Processing Magazine*, ASSP-3(1) :4–16, 1986.
- [Roach *et al.*, 1996] P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, Marasek, A. Marchal, E. Meister, & K. Vicsi. BABEL : An Eastern European Multi-language Database. In *Proceedings of ICSLP*, volume 3, pages 1892–1893, Philadelphia, 1996.
- [Salor *et al.*, 2002] O. Salor, B.L. Pellom, T. Çiloflu, & M. Demirekler. On developing new text and audio corpora and speech recognition tools for the turkish language. In *Proceedings of ICSLP*, Denver, 2002.
- [Salor *et al.*, 2006] O. Salor, T. Ciloglu, B. Pellom, & M. Demirekler. Middle East Technical University Turkish Microphone Speech v 1.0. Philadelphia, 2006. Linguistic Data Consortium.
- [Sampson, 1985] G. Sampson. *Writing Systems*. Stanford University Press, 1985.
- [Schmid, 1994] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*, Manchester, 1994.
- [Schultz & Kirchhoff, 2006] T. Schultz & K. Kirchhoff. *Multilingual Speech Processing*. Elsevier, Academic Press, 2006.
- [Schultz & Waibel, 2001] T. Schultz & A. Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35 :1-2 :31–51, 2001.
- [Schultz *et al.*, 2007] T. Schultz, A.W. Black, S. Badaskar, M. Hornyak, & J. Kominek. SPICE : Web-based Tools for Rapid Language Adaptation in Speech Processing Systems. In *Proceedings of Interspeech*, pages 2125–2128, Antwerp, 2007.
- [Schultz, 2002] T. Schultz. GlobalPhone : A Multilingual Speech and Text Database developed at Karlsruhe University. In *Proceedings of ICSLP*, Denver, 2002.
- [Seid & Gamback, 2005] H. Seid & B. Gamback. A speaker independent continuous speech recognizer for Amharic. In *Proceedings of Interspeech*, Lisboa, 2005.
- [Shannon, 1948] C.E. Shannon. A mathematical theory of communication. In *Bell System Technical Journal*, volume 27, pages 379–423 and 623–656, 1948.
- [Siivola *et al.*, 2007] V. Siivola, T. Hirsimäki, & S. Virpioja. On growing and pruning kneser-ney smoothed n-gram models. In *IEEE Transactions on Speech, Audio and Language Processing*, volume 15(5), pages 1617–1624, 2007.

- [Stolcke, 2002] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*, pages 901–904, Denver, 2002.
- [Stücker & Schultz, 2004] S. Stücker & T. Schultz. A grapheme based speech recognition system for Russian. In *Proceedings of SPECOM*, St. Petersburg, 2004.
- [The United Nations Development Program, 2006] The United Nations Development Program. *Human Development Report*. Hoechstetter Printing Co, Pittsburgh, 2006.
- [Tomokiyo, 1991] L. M. Tomokiyo. *Recognizing non-native speech : characterizing and adapting to non-native usage in speech recognition*. PhD thesis, INRIA Lorraine, 1991.
- [UNDP, 2006] UNDP. Human Development Report. Technical report, Unesco, 2006.
- [Van Eynde, F. and Gibbon, D., 2000] Van Eynde, F. and Gibbon, D. *Lexicon development for speech and language processing*. Kluwer Academic Publishers, Dordrecht, 2000.
- [Viterbi, 1967] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269, 1967.
- [Wright, 1998-2003] S. Wright. Les langues sur Internet. Technical report, Université Aston, 1998-2003.
- [Xiang *et al.*, 2006] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, & J. Makhoul. Morphological decomposition for Arabic broadcast news transcription. In *Proceedings of ICASSP*, volume I, pages 1089–1092, Toulouse, 2006.
- [Yacob, 2003] D. Yacob. Application of the Double Metaphone Algorithm to Amharic Orthography. In *Proceedings of the International Conference of Ethiopian Studies XV*, Köln, Rüdiger Köppe Verlag, 2003.
- [Young & Chase, 1998] S.J. Young & L. Chase. Speech recognition evaluation : a review of the U.S. CSR and LVCSR programmes. *Computer Speech and language*, 12(4) :263–279, 1998.
- [Young *et al.*, 1994] S. Young, J. Odell, & P. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *ARPA Workshop on Human Language Technology*, pages 286–291, 1994.
- [Young, 1993] S. Young. The HTK Hidden Markov Model Toolkit : Design and Philosophy. In *Technical Report TR.153*, Cambridge, 1993.
- [Young, 1996] S.J. Young. A review of Large-Vocabulary Continuous Speech Recognition. *IEEE Signal Processing Magazine*, 13(5) :45–57, 1996.
- [Çarkı *et al.*, 2000] K. Çarkı, P. Geutner, & T. Schultz. Turkish LVCSR : Towards better Speech Recognition for Agglutinative Languages. In *Proceedings of ICASSP*, Istanbul, 2000.