UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Reconnaissance et traduction automatique de la parole de vidéos arabes et dialectales

THÈSE

présentée et soutenue publiquement le 17 novembre 2020

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Mohamed Amine Menacer

Composition du jury

Rapporteurs :	Pr. Yannick Estève	Université Avignon, France.
	Pr. Mohand Tahar Kechadi	University College Dublin, Irlande.
Examinateurs :	Dr. Martine Adda-Decker	Université Sorbonne Nouvelle, France.
	Pr. Chiraz Latiri	Université Tunis El Manar, Tunisie.
	Pr. Kamel Smaïli	Université de Lorraine, France.
	Dr. Denis Jouvet	Université de Lorraine, France.

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Tout d'abord je remercie mon directeur de thèse Kamel Smaili qui a pris le rôle d'un enseignant, d'un ami et d'un père ne cherchant que mon succès et ma réussite. Un grand merci à mon co-directeur Denis Jouvet, sans son aide et celle de Kamel, ce travail ne serait jamais achevé. Merci de m'avoir encadré, dirigé, encouragé tout au long de ma thèse, vos conseils et remarques m'ont été d'une très grande aide. Je les remercie également pour la patience et l'attention avec laquelle ils ont lu et corrigé mon manuscrit. Pour tout cela, je tiens à leurs témoigner toute ma gratitude.

Je remercie les membres de notre équipe du projet AMIS : David Langlois, Dominique Fohr, Odile Mella, Denis Jouvet et Kamel Smaili. Cette thèse est le fruit d'une collaboration de quatre années avec eux.

Je tiens à remercier les membres de mon jury :
— Pr. Yannick Estève de l'université d'Avignon, France.
— Pr. Mohand Tahar Kechadi de university College Dublin, Irlande.
— Dr. Martine Adda-Decker de l'université Sorbonne Nouvelle, France.
— Et Pr. Chiraz Latiri de l'université Tunis El Manar, Tunisie.
malgré ces temps difficiles de maladie et de confinement, ils ont bien voulu examiner et corriger mon manuscrit ainsi qu'assister à ma soutenance dématérialisée.

Je remercie les membres de l'équipe SMArT. Tout les moments passés à travailler ensemble et à s'entraider ont fait d'eux ma seconde famille.

Je remercie mon épouse qui m'a inspiré lors de la rédaction de ma thèse et qui a pris le temps de lire et de corriger le manuscrit, merci beaucoup.

Un grand merci à mes parents, mes baux parents, mes deux soeurs Asmaa et Yasmine et mon frère Abderrahmane pour leur soutient, leur amour inconditionnel et leurs encouragements.

Enfin, je tiens à remercie mes amis et collègues de LORIA : Ameur, Sara, Fadi, Karima, Nouha, Djamel, Yacine, Valérian, Salima et Karima.

Dieu Merci

À l'être de lumière qui m'a été offert cette année, ma petite princesse Aya.
À sa mère, ma chère épouse Amina.
À ma douce mère et mon cher père.

Sommaire

Table des figures

Liste des tableaux

Introduction générale

Nous vivons dans une ère où c'est devenu primordial d'avoir accès à l'information au moment de sa diffusion. L'avance médiatique et technologique du monde nous soumet à un bombardement d'actualités à tout moment et à une vitesse phénoménale. Ces actualités peuvent avoir un impact direct ou indirect sur notre quotidien. Pour ne rien rater de ces dernières, on se retrouve fréquemment confronté à une barrière qui nous freine pour leur acquisition dans un court laps de temps : la langue. Imaginons un évènement x qui vient de se produire, et qui est présenté dans l'une des 7 117 langues existantes [1] et qu'on estime important de le connaître vu les illustrations accompagnant ce dernier, on doit trouver un moyen facile et rapide de comprendre de quoi il s'agit, c'est le but du projet AMIS, *Access to Multilingual Information and opinionS* dans lequel s'inscrivent nos travaux de thèse.

Le projet AMIS concernait le développement d'un système d'aide à la compréhension de l'information multilingue sans aucune intervention humaine permettant aux personnes de comprendre l'idée générale d'une vidéo dans une langue étrangère. Ce qu'on entend par idée générale dans ce contexte, c'est le fait de générer automatiquement un résumé de la vidéo afin d'en extraire les informations pertinentes et les traduire ensuite dans une langue compréhensible par l'utilisateur. Concrètement parlant, répondre à ce besoin implique une interaction judicieuse entre plusieurs composants, à savoir : le résumé automatique de vidéos et/ou de textes, la reconnaissance automatique de la parole et la traduction automatique. Dans le projet AMIS, on s'est focalisé sur trois langues : l'anglais, l'arabe et le français. En supposant que la langue anglaise est majoritairement compréhensible, nous avons décidé de traduire les vidéos en arabe et en français vers l'anglais.

Notre travail dans ce projet consiste à proposer une architecture pour transcrire une vidéo en texte dans une langue différente de celle utilisée dans la vidéo originale. Nous travaillons principalement sur la paire de langues arabe-anglais. Deux aspects de recherche sont abordés dans nos travaux de thèse : la reconnaissance automatique de la parole arabe et la traduction automatique de l'arabe vers l'anglais.

La reconnaissance automatique de la parole est la tâche qui consiste à analyser la voix humaine pour la transcrire sous la forme d'un texte exploitable par une machine. Ces dernières années et grâce aux nouvelles techniques d'apprentissage et à la quantité immense de données, les systèmes actuels peuvent atteindre, pour certaines langues (anglais, par exemple), un taux d'erreur comparable à celui d'un être humain. Cependant, on ne peut pas affirmer que le problème de la reconnaissance vocale soit totalement résolu. Il existe plusieurs pistes de recherche qui restent non-résolues. Certaines sont relatives à la langue (l'accent régional ou étranger, la présence de plusieurs variantes de la langue, le manque de ressources, etc) et d'autres sont relatives au signal de parole (le bruit, le chevauchement entre les locuteurs, les erreurs sémantiques, etc.).

1. source : ethnologue

Les systèmes état-de-l'art de reconnaissance automatique de la parole sont classés en deux grandes catégories selon leur architecture : les systèmes à base de plusieurs composants et les systèmes de bout en bout. Quelle que soit l'architecture du système, ce dernier vise à générer la séquence de mots $\hat{W}$ qui correspond au mieux à une séquence d'observations acoustiques O. Trouver $\hat{W}$ dans les systèmes à base de plusieurs composants est basé sur un modèle acoustique, un lexique de mots avec leurs prononciations et un modèle de langage. Du point de vue linguistique, le modèle acoustique permet de modéliser les observations acoustiques correspondant aux sons de la langue ; le modèle de langage apporte des informations sur les suites possibles de mots et enfin, le lexique définit le vocabulaire et les différentes variantes de prononciation de chaque mot. Dans les systèmes de bout en bout, ces trois composants sont remplacés par un seul modèle à base de réseaux de neurones récurrents, *Recurrent Neural Network* (RNN) [Graves *et al.*, 2006, Sutskever *et al.*, 2014].

L'avantage de ces approches est qu'elles sont indépendantes de la langue, leur mise en place dépend d'une collection de données de la langue à reconnaître, en particulier des données orales et textuelles. Néanmoins, prendre en considération les caractéristiques de certaines langues est indispensable pour booster les performances du système de reconnaissance automatique de la parole, c'est le cas de la langue arabe. Cette dernière est une langue riche en consonnes et pauvre en voyelles. De plus, l'absence de l'indication des voyelles dans le texte rend difficile leur apprentissage par le modèle acoustique. Du point de vue linguistique, la richesse morphologique de l'arabe et la simplification de l'écriture de certains mots en remplaçant une lettre par une autre augmente le nombre d'entrées dans le vocabulaire. Pour un même nombre d'entrées, le taux de mots hors vocabulaire en arabe est toujours plus élevé par rapport à l'anglais. Cela impacte le taux d'erreur vu que chaque mot hors vocabulaire ne sera pas reconnu par le système de reconnaissance de la parole et qu'il entraîne environ 1,2 erreurs dans les calcul du score final [Rosenfeld, 1995]. En partant de ces caractéristiques, nous proposons une *recette* pour développer un système de reconnaissance automatique de la parole arabe dénommé ALASR, *Arabic Loria Automatic Speech Recognition*.

L'arabe standard n'étant pas la langue maternelle des arabes, son utilisation est souvent restreinte à l'enseignement dans les écoles, aux livres, aux journaux, aux magazines et aux médias officiels. De fait, les arabes utilisent le dialecte dans leurs conversations quotidiennes. Le dialecte ou l'arabe dialectal se divise en parlers locaux différents selon les régions. Dans le cadre de cette thèse, nous travaillons sur l'adaptation du système ALASR pour le dialecte algérien. Ce dernier est l'un des dialectes maghrébins qui se caractérisent par leur difficulté à être traités par les systèmes de reconnaissance automatique de la parole. Du point de vue linguistique, cette difficulté est principalement due à l'enrichissement de son vocabulaire grâce aux nombreux emprunts faits aux autres langues, notamment le français. Du point de vue computationnel, le manque de ressources nécessaires pour apprendre les différents modèles est le problème majeur qui rend difficile la modélisation du dialecte algérien. Notre approche pour remédier à ce problème consiste à tirer profit des données provenant d'autres langues qui ont un impact sur le dialecte algérien.

Une fois la reconnaissance de la parole arabe terminée, le texte correspondant est utilisé pour la traduction automatique. Plusieurs approches ont été proposées dans la littérature pour automatiser le processus de traduction. L'approche dominante ces 20 dernières années est l'approche statistique à base de segments [Och, 1999, Koehn *et al.*, 2003]. Elle est basée sur la même idée que celle appliquée dans les systèmes de reconnaissance automatique de la parole, notamment l'utilisation de deux modèles celui de la traduction et de langage pour maximiser la probabilité de traduire une phrase source en une phrase cible. Récemment, les modèles de bout en bout à base de réseaux de neurones ont montré d'incroyables performances dans la traduction automa-

tique [Kalchbrenner et Blunsom, 2013, Sutskever *et al.*, 2014, Cho *et al.*, 2014a]. Pour certaines paires de langues, il est presque impossible de distinguer les traductions générées par l'approche neuronale et celles générées par un être humain [Wu *et al.*, 2016]. Cependant, ces approches nécessitent une quantité massive de données parallèles, de l'ordre de millions de phrases parallèles pour l'apprentissage des modèles. Or, les corpus parallèles sont coûteux à construire car ils nécessitent une expertise humaine, et ils sont souvent non disponibles pour les langues peu dotées en ressources comme les dialectes arabes. En partant de ce constat, nous menons une étude comparative entre l'approche statistique à base de segments et l'approche neuronale dans un cadre où peu de données parallèles sont disponibles. En effet, notre objectif de départ se focalisait sur la traduction du dialecte algérien. Or, les corpus parallèles dialectaux sont peu disponibles. Pour cette raison nous menons cette étude sur un corpus arabe-anglais de taille réduite. Nous étudions en particulier l'impact de quelques techniques visant à booster le modèle neuronal dans le cas où peu de données sont disponibles.

Un autre phénomène très répandu dans le langage parlé des arabophones est l'alternance codique, *code-switching*. Il se produit lorsqu'un locuteur alterne entre deux ou plusieurs langues dans un même discours. Nous menons un travail de recherche pour étudier sa présence dans les documents officiels et pour traduire des textes comprenant ce phénomène. Depuis quelques années, les données *code-switchées* sont considérées comme du bruit lors de la phase de l'apprentissage et/ou de test et elles sont écartées. Notre étude porte sur le mélange de l'arabe et de l'anglais dans un corpus parallèle extrait de documents officiels des nations unies. Nous construisons un corpus parallèle où les phrases sources sont des phrases *code-switchées* arabe-anglais et les phrases cibles sont leurs traductions dans les deux langues arabe et anglaise. En se basant sur cette ressource, nous proposons plusieurs stratégies de traduction afin de mieux traduire du texte comportant le mélange de l'arabe et de l'anglais.

Dans le cadre du projet AMIS pour la traduction de vidéos et afin de traduire la parole arabe, un système séquentiel, où la sortie du système ALASR est utilisée comme entrée du système de traduction, est proposé. Ce système soulève plusieurs problèmes dus à la propagation des erreurs entre les différents composants. Afin de minimiser ces erreurs, nous travaillons sur les points suivants : l'adaptation du vocabulaire des systèmes de reconnaissance automatique de la parole et la proposition d'une nouvelle modélisation transformant directement un signal de la parole dans une langue A en une séquence de mots dans une autre langue B.

Dans ce manuscrit de thèse et pour chaque partie de recherche, nous présentons en détail toutes nos propositions en justifiant nos choix de conception et d'adaptation des algorithmes pour la tâche de reconnaissance automatique de la parole et celle de la traduction automatique. Nous présentons également différentes expérimentations pour analyser le comportement et les performances de nos systèmes. L'organisation de ce manuscrit se présente comme suit :

Partie 1 : État de l'art

Chapitre 1 : Reconnaissance automatique de la parole. Nous présentons dans ce chapitre une vue d'ensemble sur les systèmes de reconnaissance automatique de la parole statistiques. Nous détaillons davantage le principe des modèles de Markov cachés, en particulier les modèles de mélange de gaussiennes GMM-HMM et les modèles à base de réseaux de neurones DNN-HMM tout en expliquant comment est construit un système de reconnaissance automatique de la parole. Les modèles de bout en bout ainsi que les avancées réalisées pour la reconnaissance automatique de la parole arabe et ses variantes (les dialectes) y sont présentés.

Chapitre 2 : Traduction automatique. Nous présentons dans ce chapitre les approches

proposées dans la littérature pour automatiser le processus de traduction. Nous détaillons en particulier l'approche statistique à base de segments, l'approche neuronale ainsi que la comparaison entre ces approches. Nous y présentons également les travaux réalisés dans le cadre de la traduction de/vers la langue arabe.

Partie 2 : Contributions

Chapitre 3 : ALASR : un système de reconnaissance automatique de la parole arabe. Ce chapitre est consacré pour la description de notre *recette* pour le développement du système *ALASR* : un système de reconnaissance automatique de la parole pour la langue arabe. Nous présentons, dans un premier temps, les aspects acoustiques et syntaxiques caractérisant l'arabe ainsi que les défis à surmonter pour reconnaître la parole arabe. Ensuite, nous détaillons les approches proposées pour surmonter ces défis. Enfin, nous présentons les différentes expérimentations et tests que nous avons menés pour évaluer notre système et le comparer avec les systèmes état-de-l'art.

Chapitre 4 : L'adaptation d'ALASR pour le dialecte Algérien. Nous présentons dans ce chapitre nos travaux pour adapter le système ALASR afin de reconnaître une variante du dialecte algérien : le dialecte algérois [2]. Nous commençons par présenter les aspects acoustiques et syntaxiques caractérisant ce dialecte. Nous présentons ensuite nos propositions d'adaptation des différents modèles (acoustique et linguistique). Nous discutons enfin les résultats obtenus tout en les comparant avec les résultats état-de-l'art.

Chapitre 5 : Approche statistique vs. neuronale dans la traduction automatique. Une étude comparative entre l'approche statistique à base de segments et l'approche neuronale pour la traduction automatique dans le cadre où peu de données sont disponibles est présentée dans ce chapitre.

Chapitre 6 : Traduction automatique du texte *code-switché*. Nous présentons dans ce chapitre le phénomène du *code-switching* dans la langue arabe tout en se concentrant sur les passages entre l'arabe et l'anglais. Nous détaillons notre approche pour construire un corpus parallèle *code-switché* avec deux traductions de référence ainsi que les techniques proposées pour traduire du texte *code-switché*.

Chapitre 7 : Projet AMIS et contributions. Ce dernier chapitre a pour objectif de résumer l'adaptation de nos travaux au contexte du projet AMIS. Nous présentons nos approches pour la traduction de la parole arabe vers l'anglais : l'approche basée sur un modèle séquentiel et l'approche basée sur les modèles de bout en bout. Nous présentons en particulier nos techniques pour l'adaptation des vocabulaires des systèmes de reconnaissance automatique de la parole afin de réduire la propagation des erreurs dans le modèle séquentiel. Nous clôturons par une description de quelques pistes de recherche afin d'améliorer nos approches de traduction automatique de la parole arabe.

2. Dans le reste du manuscrit, on utilise le dialecte algérien pour désigner le dialecte algérois.

Première partie

État de l'art

Chapitre 1

Reconnaissance automatique de la parole

Sommaire

La parole est le moyen de communication le plus utilisé par les êtres humains. Avec une vitesse de prononciation qui ne dépasse pas les 150 mots par minute en comparaison avec la vitesse d'écriture qui peut atteindre 50 mots par minute, les personnes préfèrent utiliser leur voix au lieu du texte pour communiquer. Toutefois, ce n'est pas seulement une question de vitesse, plusieurs facteurs sont importants quand il s'agit de la parole : le contexte, le lieu, les personnes avec qui on parle, les idées sous-entendues véhiculées par la parole, etc. Tous ces facteurs rendent ce moyen de communication très complexe à automatiser.

Au cours de ces dernières années, des progrès importants ont été réalisés concernant les performances de reconnaissance automatique de la parole. Pour certaines langues, la langue anglaise par exemple, on peut atteindre des taux proches de ceux d'un être humain, et ceci grâce, d'une part, aux nouvelles techniques d'apprentissage et, d'autre part, à la quantité immense de données utilisées pour développer ce genre de système.

Ce chapitre commence par la présentation d'une vue d'ensemble sur les systèmes de reconnaissance automatique de la parole. Nous abordons, par la suite, les techniques état de l'art pour apprendre les différents modèles tout en se focalisant sur les éléments sur lesquels nous sommes intervenus. Le chapitre se termine par la présentation des avancées réalisées pour la reconnaissance automatique de la parole arabe ainsi que ses variantes (les dialectes).

1.1 Reconnaissance automatique de la parole : du signal au langage

Les premiers systèmes de reconnaissance automatique de la parole datent des années soixante. Il s'agissait de systèmes avec un vocabulaire très restreint et qui ne reconnaissaient que des mots isolés. La première modélisation statistique sur laquelle se base les systèmes actuels a été proposée par [Jelinek, 1976]. Elle s'inscrit dans le cadre des approches basées sur le modèle du canal bruité introduit par [Shannon, 1948] et largement utilisé dans plusieurs domaines (traduction automatique, résumé automatique de texte, correction automatique d'orthographe, etc.). Ce modèle comporte une source d'information (la parole), un canal de transmission (la production de parole, l'analyse acoustique et la transmission) et un décodeur (le décodage linguistique). Les systèmes actuels de reconnaissance automatique de la parole comportent plusieurs composants comme illustré dans la figure 1.1.

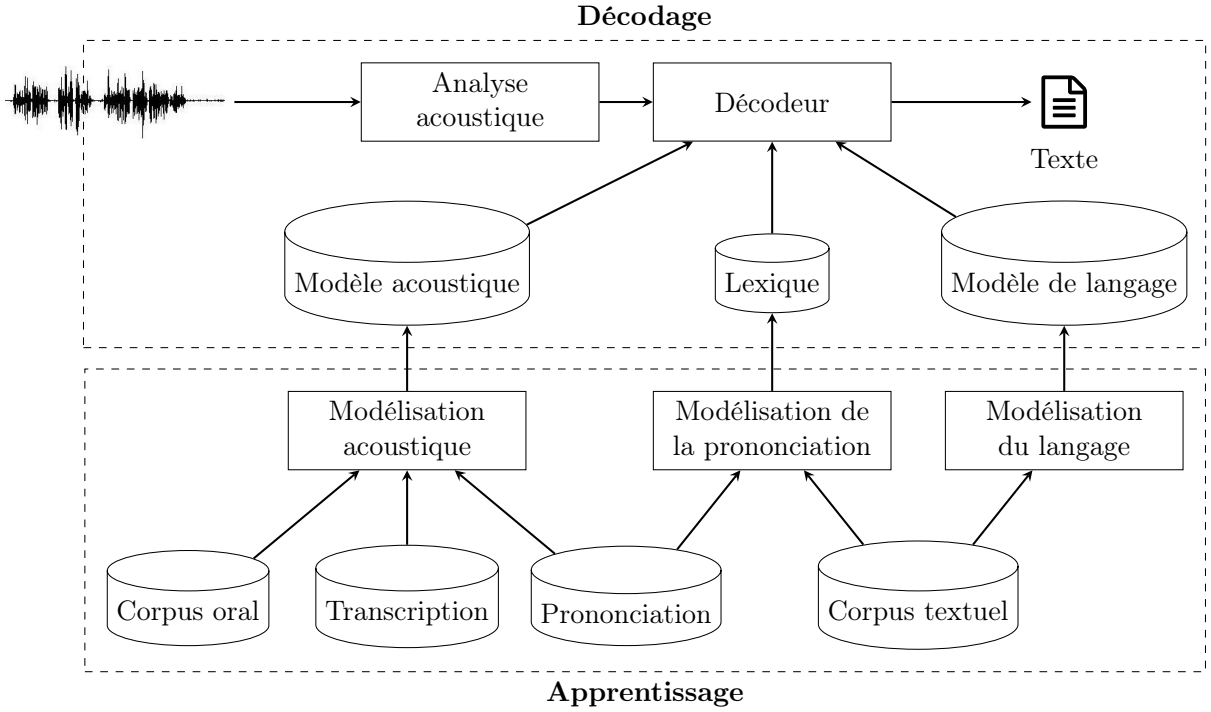


FIGURE 1.1 – Les composants du système de reconnaissance automatique de la parole basés sur l'approche statistique.

En pratique, le développement d'un système de reconnaissance automatique de la parole est basé sur deux processus principaux : l'apprentissage et le décodage. Lors de l'apprentissage, les différents paramètres de différents modèles sont estimés en utilisant des données orales pour le modèle acoustique et des données textuelles pour le modèle de langage. Le processus de décodage consiste à trouver pour un signal acoustique le texte qui lui correspond au mieux. Pour accomplir cette tâche le signal analogique doit être numérisé et converti pour extraire des observations acoustiques encapsulant le message linguistique. Le décodeur génère par la suite la séquence de mots qui correspond à ces observations en se basant sur le modèle acoustique, un lexique de mots avec leurs prononciations, et un modèle de langage. Du point de vue linguistique, le modèle acoustique permet de modéliser les observations acoustiques correspondant aux sons de la langue,

le modèle de langage apporte des informations sur les suites possibles de mots et enfin, le lexique définit le vocabulaire et les différentes variantes de prononciation de chaque mot.

D'une manière plus formelle, le processus de la reconnaissance automatique de la parole consiste à trouver pour une observation acoustique O, la séquence de mots $\hat{W}$ qui maximise la probabilité conditionnelle $P(W|O)$ comme le montre l'équation 1.1

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{1.1}$$

Le terme $P(O|W)$ correspond à la probabilité donnée par le modèle acoustique et $P(W)$ est la probabilité de la séquence de mots W calculée par le modèle de langage. Le terme $P(O)$ est la probabilité à priori de l'observation acoustique qui est indépendante du modèle de langage et du modèle acoustique et qui peut être ignorée lors de l'opération de maximisation. Cela conduit à l'équation fondamentale sur laquelle se base les systèmes statistiques de reconnaissance automatique de la parole :

$$\hat{W} = \arg\max_{W} P(O|W)P(W) \tag{1.2}$$

Dans ce qui suit, chaque composant du système de reconnaissance automatique de la parole sera présenté en détail.

1.1.1 Traitement du signal

Le signal de la parole comporte de nombreuses informations autres que le message linguistique comme des informations sur le locuteur, sur la langue, sur les conditions d'enregistrement, etc. Ces informations ne sont pas forcément pertinentes lors du décodage et elles sont des sources de variabilité du signal de parole. Afin d'exploiter au mieux le signal de la parole, il est nécessaire de le numériser et d'en extraire seulement les paramètres relatifs au message linguistique. Ces paramètres vont servir après comme paramètres d'entrée pour l'algorithme de reconnaissance.

Le signal de la parole est un signal analogique, sa numérisation consiste à observer l'évolution de l'amplitude du signal et de produire en sortie une représentation de cette valeur d'amplitude sous forme d'un mot binaire codé sur un ou plusieurs octets. Ce processus comporte deux étapes principales : l'échantillonnage et la quantification.

Échantillonnage il consiste à prélever des valeurs à une fréquence bien définie. Le théorème de Shannon [Shannon, 1948] impose d'échantillonner le signal à une fréquence égale au moins au double de la fréquence maximale contenue dans le signal. Pour le signal de la parole, dont les fréquences peuvent atteindre 8 kHz, le signal d'entrée doit être échantillonner au minimum 16000 fois par seconde, c'est-à-dire à une fréquence d'échantillonnage de 16 kHz.

Quantification les valeurs prélevées du signal doivent être converties en valeurs binaires sur un certain nombre de bits. Cette étape va inévitablement ajouter une erreur de mesure, appelée le bruit de quantification, inhérente à la compression des données analogiques vers un format dont la résolution est limitée par le nombre de bits. Pour des enregistrements de parole, une résolution de 16 bits entraîne généralement une erreur de quantification négligeable et sans conséquence sur les performances des systèmes de reconnaissance automatique de la parole.

Le signal numérisé est converti dans un second temps en des paramètres acoustiques : un ensemble de vecteurs qui sont plus pertinents pour le traitement de la parole. Vu la non-stationnarité

du signal de la parole, ces paramètres acoustiques ne peuvent pas être estimés globalement. Ils sont estimés sur des trames, c'est-à-dire sur des segments de 10 à 30 sur lesquels le signal est jugé stationnaire. Cette segmentation en trames à courte durée produit des discontinuités aux frontières des trames ce qui conduit, après, à des lobes secondaires dans le domaine spectral. Ces discontinuités sont réduites via l'application d'une fenêtre de Hamming.

Pour chaque segment de parole (trame), le signal obtenu après application de la fenêtre de Hamming est converti en un vecteur acoustique en utilisant des traitements qui sont principalement mis en œuvre dans le domaine spectral. Différentes techniques de calcul de coefficients acoustiques existent. Parmi ces techniques, on peut citer : *Linear Predictive Coding* (LPC) [O'Shaughnessy, 1988], *Perceptual Linear Prediction* (PLP) [Hermansky, 1990], et la technique la plus utilisée dans les systèmes actuels *Mel Frequency Cepstral Coefficients* (MFCC). Les paramètres MFCC peuvent être calculés selon le processus décrit ci-dessous (figure 1.2) [Davis et Mermelstein, 1980].

FIGURE 1.2 – Processus d'extraction des paramètres MFCC.

La transformée de Fourier est appliquée pour chaque trame fenêtrée afin de passer du domaine temporel au domaine fréquentiel. Ce passage permet de calculer l'énergie pour les différentes fréquences ce qui simule le travail effectué par la cochlée (un organe de l'oreille interne). En effet, la cochlée vibre à différents endroits selon les fréquences perçues pour informer le cerveau que certaines fréquences sont présentes. Cette transformation permet, ainsi, une nouvelle représentation du signal qui est appelée le spectrogramme. Il s'agit d'une représentation tridimensionnelle qui indique la répartition énergétique du son en fonction du temps et des fréquences (voir figure 1.3). Cette représentation met en évidence l'évolution de l'énergie dans le temps, ainsi que la position fréquentielle des zones énergétiques ; ces informations permettent d'identifier les sons (phonèmes) prononcés. En pratique, les modèles acoustiques modélisent les formes spectrales de la figure 1.3 observés pour chaque unité acoustique [3].

L'oreille perçoit des écarts de fréquences faibles dans les basses fréquences, mais les écarts doivent être plus larges dans les hautes fréquences pour être perçues. Pour reproduire ce phénomène on utilise des bancs de filtres dont la largeur augmente. Ceci est achevé par des bancs de filtres sur une échelle de Mel (on utilise en pratique 25 filtres). Le premier filtre est très étroit et donne une indication de la quantité d'énergie disponible dans les très basses fréquences. À mesure que les fréquences augmentent, les filtres s'élargissent.

Le logarithme de l'énergie dans chaque bande de fréquence (filtre Mel) est calculé par la suite. Cela est également motivé par le système d'auditif humain : on n'entend pas l'intensité sonore sur une échelle linéaire.

La dernière étape consiste à calculer la transformée discrète en cosinus pour décorréler les bandes d'énergie calculées avec les bancs de filtres de Mel [Davis et Mermelstein, 1980]. En pratique, seulement les treize premiers coefficients sont conservés ; les coefficients plus élevés correspondent à des détails fins sur la forme du spectre, et donc pas nécessairement pertinents pour la reconnaissance de la parole.

3. Les unités acoustiques correspondent typiquement aux sons élémentaires de la langue (phonèmes), ainsi que les silences et les bruits.

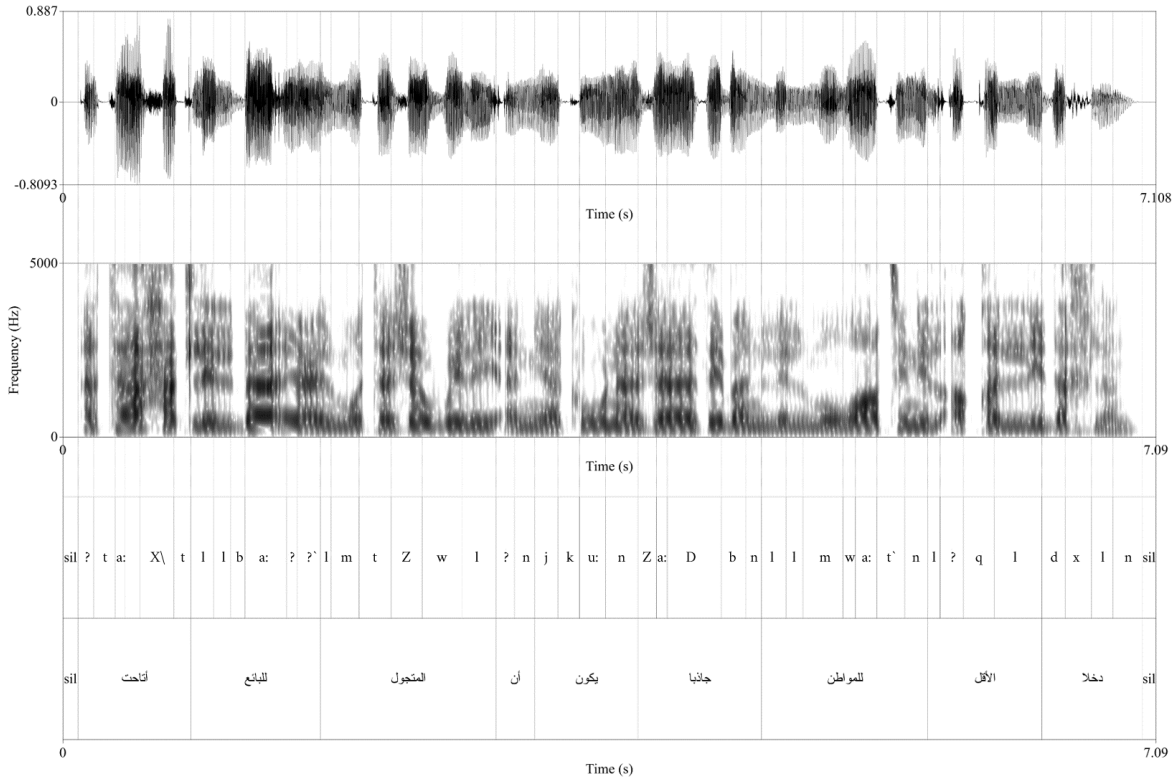


FIGURE 1.3 – Exemple d'alignement entre les unités acoustiques (phonèmes arabes avec codage SAMPA) et le spectrogramme du signal qui correspond à la prononciation d'une phrase arabe.

1.1.2 Modélisation acoustique

La fonction du modèle acoustique est d'estimer la probabilité $P(O|W)$ de l'équation 1.2. Le développement d'un tel modèle est l'un des principaux défis de la reconnaissance automatique de la parole. Cela est principalement dû à la variabilité des observations acoustiques qui est principalement due au locuteur (variabilité intra et inter locuteur, état émotionnel, spontanéité, etc.), au bruit ambiant et prise de son et au canal de transmission. Ajoutons à cela la coarticulation entre son qui impacte la prononciation des phonèmes selon leur contexte. Le même constat est également valable en ce qui concerne la coatriculation entre mots où leur prononciation peut changer en fonction des mots voisins (par exemple la notion de liaison en français). Un bon modèle acoustique doit prendre en considération tous ces facteurs pour que le système de reconnaissance automatique de la parole soit robuste.

La première étape pour développer le modèle acoustique est le choix des unités acoustiques qui peuvent être soit des mots, des syllabes, des caractères ou des phonèmes. Le choix de ces unités dépend du domaine d'application. Dans le cas où le vocabulaire est restreint, un modèle à base de mots donne de bons résultats à condition que chaque mot ait suffisamment d'exemples dans le corpus d'apprentissage. En revanche, dans le cas de la parole continue où le vocabulaire est plus riche, chaque mot peut avoir plus qu'une prononciation, et un bon modèle acoustique doit prendre en considération toutes les prononciations possibles selon le contexte de chaque mot. Cela nécessite une grande quantité de données pour apprendre les différentes variantes de prononciation. Pour remédier à ce problème, les modèles à base d'unités acoustiques plus

petites (syllabes ou phonèmes) sont plus appropriés pour ce cas de figure. Les modèles à base de phonèmes sont les plus utilisés dans les systèmes de reconnaissance automatique de la parole actuels. Ils nécessitent une quantité modérée de données pour estimer de manière robuste les modèles acoustiques de chaque phonème.

Il existe plusieurs approches pour la modélisation acoustique. Ces approches sont basées sur des modèles statistiques dont l'apprentissage s'effectue sur des corpus acoustiques/oraux : une collection de paroles ainsi que leurs transcriptions correspondantes. Dans ce qui suit, on détaillera quelques-unes des approches les plus utilisées, dont : les approches basées sur les modèles de Markov caché, *Hidden Markov Model* (HMM), et les modèles basés sur les réseaux de neurones (DNN-HMM et les modèles de bout en bout (end2end)). Ce sont les approches utilisées dans nos travaux de thèse.

Approches à base du modèle de Markov caché

La chaîne de Markov est un processus stochastique représenté par un ensemble d'états $S_i,\ 1 \leq i \leq N$ (avec $N \in \mathbb{N}$) dont chacun dépend du précédent et un ensemble de transitions. Une transition représente la possibilité de passer d'un état à un autre. La figure 1.4 illustre un modèle de Markov à trois états ($N = 3$). Les transitions sont unidirectionnelles et tous les états ont des transitions vers tous les autres états, y compris vers eux-mêmes. Chaque transition a une probabilité d'être empruntée ($a_{ij},\ i,j \in \{1,2,3\}$) et cette probabilité peut éventuellement être nulle. La somme des probabilités des transitions partant d'un état est toujours égale à 1.

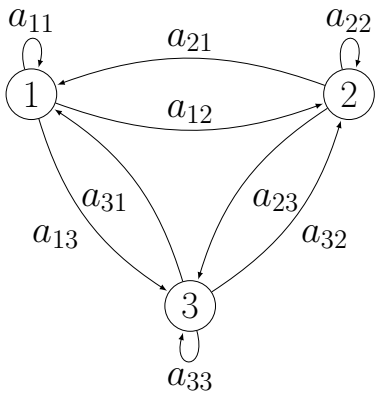


FIGURE 1.4 – Chaîne de Markov à trois états.

Le modèle de Markov est souvent associé à la notion du temps. À chaque unité de temps, on opère une transition d'un état à un autre, ce qui génère finalement une séquence d'états avec sa probabilité de se produire. Cela signifie que la séquence d'états est directement observable au fil du temps contrairement aux modèles de Markov cachés. Dans ces derniers, chaque état émet des observations qui, elles, sont observables. On ne travaille donc pas sur la séquence d'états, mais sur la séquence d'observations générées par les états. En appliquant ce principe sur la modélisation acoustique de la parole, chaque unité acoustique (phonème le plus souvent) est représentée par un ensemble de trois états, les observations émises par chaque état représentent les vecteurs MFCC calculés lors de l'extraction des paramètres acoustiques. Cette modélisation est schématisée par la figure 1.5.

Les deux états S_{init} and S_{fin} sont des états non-émetteurs, ils sont utilisés uniquement pour simplifier le processus de concaténation des modèles de phonèmes pour la génération des modèles associés aux mots.

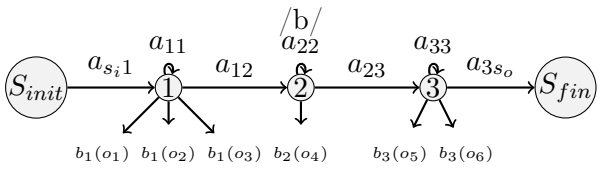


FIGURE 1.5 – Modèle de Markov caché à trois états pour le phonème /**b**/.

Dans un modèle de Markov caché, à chaque instant t, une transition de l'état actuel i à l'un de ses états connectés j est exécutée avec une probabilité a_{ij}. Chaque état i peut émettre une observation acoustique o_t avec une probabilité $b_i(o_t)$. Mathématiquement, un modèle de Markov caché se décrit par les paramètres suivants :

— Le nombre N des états du modèle.
— Les probabilités de transition d'états a_{ij}. Ces probabilités sont représentées par une matrice A de taille $N \times N$.
— La densité de probabilité d'observation associée à l'état i, c'est-à-dire l'estimation des probabilités $(b_i(o_t)) = P(o_t|S_i)$ avec $1 \leq i \leq N$.
— La distribution de probabilité d'être à un état à l'instant initial, $\Pi = (\pi_i)$ avec $1 \leq i \leq N$.

En reconnaissance automatique de la parole, la probabilité de produire une observation acoustique par chaque état est modélisée par une mixture de lois gaussiennes, d'où les modèles de Mélange de gaussiennes, *Gaussian Mixture Model-Hidden Markov Model* (GMM-HMM). Cela signifie que les probabilités d'émission $b_i(o_t)$ du modèle de Markov caché sont estimées par une somme de M lois normales :

$$b_i(o_t) = \sum_{m=1}^{M} c_{im} \mathcal{N}(o_t; \mu_{(im)}, \Sigma_{(im)}) \tag{1.3}$$

le terme c_{im} est un coefficient de pondération associé à chaque composante de loi normale. Les $\mu_{(im)}$ et $\Sigma_{(im)}$ sont respectivement les vecteurs moyennes et les matrices de covariance de chaque lois normale. Nous donnons dans la figure 1.6 un exemple de distribution de probabilité modélisant la génération d'une observation par un état donné en utilisant une seule loi normale (courbe à gauche) et une combinaison de trois distributions de loi gaussienne avec des poids différents (courbe à droite). Dans cet exemple, nous avons supposé que chaque observation acoustique est de dimension une au lieu de 13 (la dimension des observations MFCC) pour simplifier la visualisation des distributions.

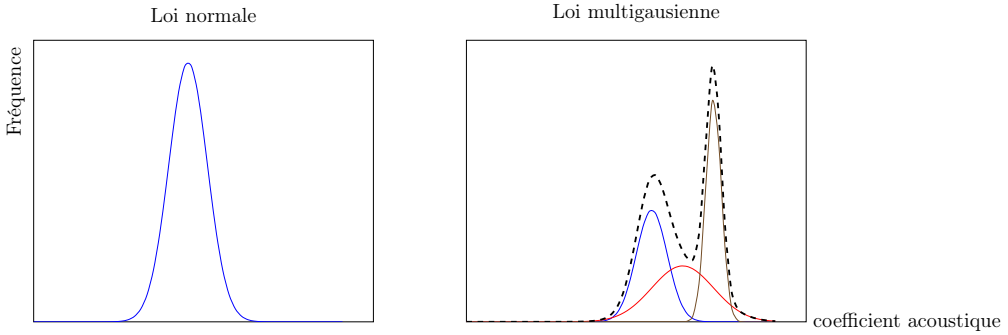


FIGURE 1.6 – Exemple de distribution de probabilité d'émission $b_i(o_t)$.

Un modèle de Markov caché est ainsi décrit par le jeu de paramètres $\lambda = (A, \mu, \Sigma, \Pi)$ (avec μ et Σ sont les paramètres des lois gaussiennes) sont estimés lors de la phase d'apprentissage. Ce modèle est typiquement utilisé pour générer une séquence d'observation $O = o_1 o_2 \dots o_T$ (avec T est le nombre total des observations à générer) ou pour estimer la probabilité qu'une séquence d'observations soit générée par ce modèle. Pour pouvoir utiliser et mettre en pratique les modèles de Markov cachés, on doit répondre aux trois problèmes suivants [Rabiner, 1989] :

Apprentissage. Étant donné un modèle de Markov caché de N états et une séquence d'observations de longueur T ($O = o_1 o_2 \dots o_T$) ; lors de l'apprentissage, on estime les paramètres λ du modèle qui maximisent la probabilité d'apparition de notre séquence d'observations $P(O|\lambda)$. De nos jours, il n'existe aucun algorithme permettant de résoudre ce problème ; il existe néanmoins un algorithme permettant d'approcher la solution, il s'agit de l'algorithme de *Baum-Welch* [Baum et Eagon, 1967]. C'est un algorithme itératif qui est un cas particulier de l'algorithme d'Estimation-Maximisation (EM) dont l'idée est la suivante :

1. Initialiser le modèle avec des probabilités quasiment équiprobables ou des approximations correspondantes à nos attentes.
2. Réévaluer le modèle par rapport à la séquence d'observations O utilisée pour l'apprentissage, pour obtenir un nouveau modèle $\bar{\lambda} = (\bar{A}, \bar{\mu}, \bar{\Sigma}, \bar{\Pi})$;
3. Calculer la probabilité d'observer notre séquence O avec les paramètres du nouveau modèle, $P(O|\bar{\lambda})$.
4. Si $P(O|\lambda) \leq P(O|\bar{\lambda})$ alors recommencer à partir de l'étape 2.

Évaluation. Étant donné un modèle dont les paramètres sont déjà estimés $\lambda = (A, \mu, \Sigma, \Pi)$ et une séquence d'observation $O = o_1 o_2 \dots o_T$, le problème d'évaluation consiste à calculer la probabilité $P(O|\lambda)$. Cette probabilité se calcule en additionnant les probabilités de générer cette observation O pour chacune des séquences d'états possibles. L'algorithme *forward-backward*[4] [Baum et Eagon, 1967] donne la solution exacte avec une meilleure complexité $O(N^2 T)$.

Décodage. Ce dernier problème consiste à trouver pour une séquence d'observations $O = o_1 o_2 \dots o_T$, la meilleure séquence d'états génératrice. C'est la question à laquelle nous souhaitons répondre dans le cadre de la reconnaissance automatique de la parole. Contrairement aux deux problèmes précédents où la solution était unique, il existe plusieurs solutions possibles pour ce problème. En effet, chaque séquence d'observations peut avoir plusieurs séquences d'états génératrices, donc, il faut choisir parmi ces séquences candidates la séquence la plus probable. En pratique, l'algorithme de *Viterbi* [Forney, 1973] basé sur la programmation dynamique, est utilisé pour trouver efficacement la meilleure séquence d'états.

Pour un mot w, le modèle acoustique correspondant est fabriqué en concaténant les modèles des phonèmes en fonction de la prononciation spécifiée dans le dictionnaire de prononciations. Cependant, modéliser chaque mot en une séquence de phonèmes indépendants ne parvient pas à capturer la grande variabilité de prononciation qui existe dans la parole. Ces modèles indépendants sont appelés des modèles *monophones*. Pour tenir compte de la coarticulation, au lieu de modéliser chaque phonème par un modèle de Markov caché à trois états, on utilise ce dernier pour modéliser le phonème dans son contexte gauche et droit, c'est ce qu'on appelle les modèles *triphones*. La figure 1.7 illustre la concaténation de trois modèles de Markov cachés pour

4. Pour estimer la probabilité $P(O|\lambda)$, seulement la partie *forward* de l'algorithme *forward-backward* est nécessaire. La partie *backward* est utilisée dans l'algorithme *Baum-Welch* pour l'estimation de paramètres du modèle.

reconnaître le mot arabe بَاب *bāb* (porte). Ce mot est constitué de trois phonèmes /b/, /a: / et /b/ [5] dont chacun est modélisé par un modèle à trois états dans le modèle *monophones*. Dans le modèle *triphones*, chaque modèle de Markov caché modélise un phonème dans un contexte gauche et droit donné.

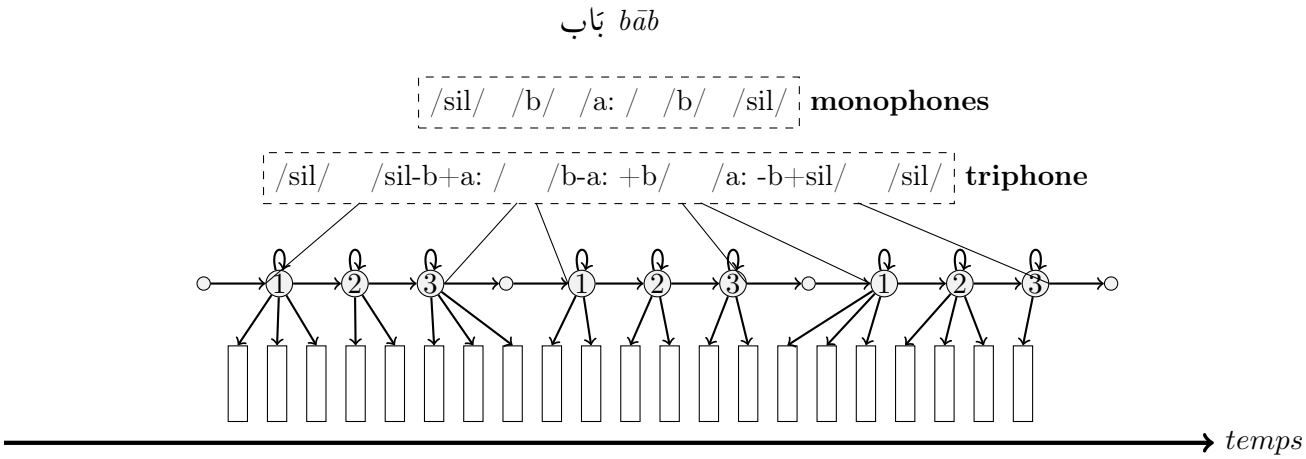


FIGURE 1.7 – Concaténation des modèles de Markov cachés à trois états pour modéliser le mot arabe بَاب *bāb* (porte).

La solution de modéliser chaque *triphone* par un HMM à trois états est très coûteuse. Prenons l'exemple de l'arabe, le sujet de nos travaux de thèse. En arabe, il existe 38 phonèmes selon le système de transcription phonétique SAMPA *(Speech Assessment Methods Phonetic Alphabet* [6]*)*. Cela implique qu'il existe $38^3 = 54872$ *triphones* potentiels. Apprendre tous ces *triphones* nécessite une grande quantité de données où chaque combinaison doit avoir suffisamment d'exemples dans le corpus d'apprentissage, ce qui n'est pas faisable en pratique. La solution est de partager des paramètres entre des *triphones* similaires pour améliorer la fiabilité de la modélisation. Ce processus est réalisé en se basant sur les arbres de décision [Young *et al.*, 1994]. Un arbre de décision binaire est associé à chaque état où les nœuds de cet arbre sont construits en se basant sur des connaissances linguistiques. Le point de départ est de rassembler tous les états modélisant le phonème courant dans une seule classe au niveau du nœud racine de l'arbre de décision. La classe au niveau de chaque nœud est subdivisée successivement en posant une question linguistique binaire qui porte sur le contexte phonétique gauche ou droit du phonème pris en compte. Les états au niveau des feuilles sont liés pour former un nouveau modèle avec un nombre réduit d'états. Dans l'exemple de la figure 1.8, nous associons la question suivante au noeud racine de l'arbre de décision : "est-ce que le contexte droit des *triphones* qui partagent le /a/ comme phonème central est un /b/ ?"

5. On utilise le système phonétique de *Speech Assessment Methods Phonetic Alphabet* (SAMPA).
6. Voir : https://www.phon.ucl.ac.uk/home/sampa/arabic.htm

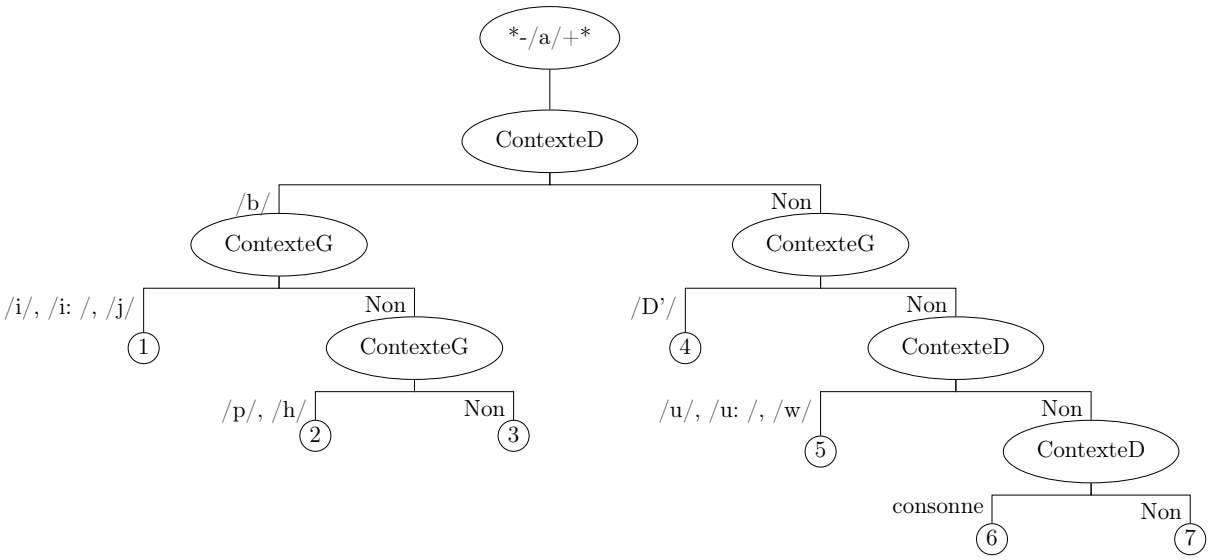


FIGURE 1.8 – Exemple d'un arbre de décision pour les *triphones* qui ont le /a/ comme phonème central.

En posant des questions sur le contexte gauche et droit du phonème /a/, les états fournissant la même réponse sur toutes les questions de l'arbre, partageront les même paramètres.

À partir de l'an 2012, grâce aux nouvelles performances des machines, les réseaux de neurones ont été adaptés et mis en pratique pour la modélisation acoustique afin d'améliorer les modèles GMM-HMM [Fohr *et al.*, 2017, Hinton *et al.*, 2012a]. Ces modèles sont utilisés pour remplacer les densités multigaussiennes des modèles GMM-HMM.

Approches neuronales

La modélisation acoustique basée sur les réseaux de neurones est vue comme un problème de classification. En effet, les classes représentent l'ensemble des états du modèle de Markov caché, et à chaque instant t, on cherche a classifier l'observation acoustique correspondant à cet instant. Ces modèles, connus sous le nom DNN-HMM, sont plus efficaces que les modèles GMM-HMM et c'est eux qui sont utilisés dans nos travaux de thèse.

D'une manière plus formelle, pour une observation acoustique o_t à l'instant t, la probabilité $y_t(s)$ estimée par le réseau de neurones pour l'état s du modèle de Markov caché est calculée avec la fonction $softmax$ comme suit :

$$y_t(s) = P(s|o_t) = \frac{\exp(a(s))}{\sum_{s'} \exp(a(s'))} \tag{1.4}$$

avec $a(x)$ est la fonction d'activation de la dernière couche. Les fonctions d'activation sont généralement des fonctions non linéaires utilisées pour permettre le passage ou non de l'information aux autres neurones en appliquant une transformation sur la sortie de chaque neurone. Cette dernière est représentée par la somme pondérée des sorties de neurones de la couche précédente sur laquelle on ajoute une unité de biais b. Si on considère p neurones de la couche précédente, la sortie de chaque neurone de la couche actuelle est calculée comme suit :

$$b + \sum_{i=1}^{p} w_i x_i \tag{1.5}$$

Apprendre les paramètres du réseau de neurones revient à estimer une matrice W représentant tous les poids et un vecteur de biais B.

L'un des premières architectures appliquée pour la modélisation acoustique est le perceptron multicouche. Un perceptron multicouche est un réseau de neurones organisé en plusieurs couches où l'information est propagée de la première couche, celle de l'entrée, jusqu'à la dernière, celle de la sortie ; on parle du réseau neuronal à propagation avant, *feedforward*.

Pour apprendre les paramètres du réseau de neurones, une fonction objective $\mathcal{F}$ doit être définie. Cette fonction est utilisée pour mesurer à quel point notre modèle modélise l'ensemble d'apprentissage en mesurant l'erreur entre la prédiction du réseau de neurones et la sortie attendue. Dans les problèmes de classification, il est courant d'utiliser l'opposé de la log-vraisemblance, qui est une entropie-croisée, comme fonction objective dont la formule est la suivante :

$$\mathcal{F}_{(EC)} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log y_t(s) \tag{1.6}$$

avec U est le nombre total de segments dans le corpus d'apprentissage, et T_u est le nombre de trames dans chaque segment u. Cela revient à calculer l'entropie-croisée entre la distribution représentés par les étiquettes de référence [7] et la distribution prédite par notre réseau de neurones $y_t(s)$.

L'entropie-croisée est appliquée au niveau des trames, en pratique, le modèle acoustique peut être amélioré en appliquant d'autres critères d'optimisation discriminants [Veselý *et al.*, 2013] et qui fonctionnent au niveau des séquences de mots. Ce critère vise à maximiser la probabilité de la séquence de mots qui correspond à une observation acoustique. En revanche, il n'introduit à aucun moment un lien entre la séquence prédite par le modèle W et la référence W_r, et donc il n'a pas une relation directe avec le *Word Error Rate* (WER). Ce problème est résolu en intégrant une fonction appelée la fonction de perte $L(W, W_r)$ dans le critère d'apprentissage, c'est ce qu'on appelle le *Minimum Bayes Risk (MBR)* [Gibson et Hain, 2006]. La fonction $L(W, W_r)$ peut être fixée selon plusieurs manières :

- Si elle mesure la distance de Levenshtein entre les mots des deux séquences W et W_r alors elle est identique au WER.
- Si elle mesure la distance entre les phonèmes des deux séquences W et W_r alors on parle de l'apprentissage basé le critère *Minimum phone error (MPE)* [Povey et Woodland, 2002].
- Si elle mesure la distance entre les densités, on parle de l'apprentissage basé sur le critère *state MBR (sMBR)* [Gibson et Hain, 2006, Veselý *et al.*, 2013].

Étant donné la fonction objective $\mathcal{F}$, les paramètres du réseau de neurones sont estimés avec la rétropropagation du gradient, *backpropagation algorithm* [Rumelhart *et al.*, 1986]. Cette technique est basée sur l'utilisation du gradient δ dans un processus itératif pour minimiser la fonction objective. Pour chaque itération, le poids w_i pour un neurone donné est mis à jour en se déplaçant dans la direction opposée au gradient selon la formule suivante :

$$w_i^t = w_i^{t-1} - \epsilon \frac{\delta \mathcal{F}}{\delta w_i} \tag{1.7}$$

Le terme ϵ est un hyperparamètre du réseau de neurones, appelé la vitesse d'apprentissage, *learning rate*. Le choix de cet hyperparamètre affecte le processus d'apprentissage, si sa valeur est grande alors on risque de s'éloigner de la valeur optimale du poids w_i d'où la divergence de l'algorithme. À l'inverse, si sa valeur est trop petite, alors le temps de convergence de l'algorithme sera plus long. Il existe dans la littérature plusieurs techniques d'optimisation des

7. Ces étiquettes sont obtenues grâce à l'alignement forcé exécuté par les modèles GMM-HMM.

paramètres basées sur l'algorithme de gradient et dont le choix de ϵ dépend de la technique utilisée. Parmi ces techniques on cite : la descente de gradient stochastique, *stochastic gradient descent (SGD)* [Bottou, 1998], *momentum* [Rumelhart *et al.*, 1986], *Nesterov accelerated gradient (NAG)* [Nesterov, 1983], *Adagrad* [Duchi *et al.*, 2011], *Adadelta* [Zeiler, 2012], *adaptive moment estimation (Adam)* [Kingma et Ba, 2014], *natural gradient descent* [Amari, 1998], et *RMSprop* [Hinton *et al.*, 2012b]. Nous présentons dans le chapitre 3 les différentes architectures utilisées dans le cadre de nos travaux de thèse.

1.1.3 Modélisation de la prononciation

Le vocabulaire ou le lexique de prononciations contient tous les mots qui sont reconnus par le système ainsi que leurs variantes de prononciation. Les mots sont généralement sélectionnés à partir des données textuelles. Le processus de sélection dépend fortement du domaine d'application. Ce processus a un impact important sur les performances du système de reconnaissance ; choisir un vaste vocabulaire non seulement rend le processus du décodage lent, mais nuit également à la précision de la reconnaissance. D'autre part, choisir un vocabulaire restreint mène à un taux de mots hors vocabulaire important sachant que chaque mot hors vocabulaire entraîne environ 1,2 erreurs [Rosenfeld, 1995] et ce taux est plus important dans certaines langues morphologiquement riche comme la langue arabe.

Plusieurs techniques de sélection du vocabulaire ont été proposées dans la littérature. La technique la plus simple est de choisir les N mots les plus fréquents à partir du corpus textuel. Dans le cas où plusieurs ressources textuelles sont disponibles, un poids peut être attribué à chaque ressource pour sélectionner ensuite les mots en fonction de leur fréquence d'apparition et du poids de chaque ressource. Pour ce faire, les mots doivent être classés en fonction d'un poids qui définit leur importance pour le domaine d'application. Ce poids $C(w_i)$ est calculé dans le cas de k documents selon l'équation 1.8.

$$C(w_i) = \sum_k \lambda_k C_k(w_i) \tag{1.8}$$

avec $C_k(w_i)$ est la fréquence du mot w_i dans le document k, et λ_k est le poids de chaque document. Ils existent plusieurs approches pour définir ces poids. Une approche simple consiste à utiliser la taille du texte comme un poids de chaque document dans l'interpolation linéaire. Une autre approche consiste à optimiser ces poids sur un corpus de validation en utilisant l'algorithme EM [Boisvert, 2006, Venkataraman et Wang, 2003].

La sélection des mots est une première étape pour la génération du lexique. Pour chaque mot sélectionné, les différentes variantes de prononciation doivent être restituées. Pour les langues avec suffisamment de ressources (la langue anglaise par exemple), il existe plusieurs dictionnaires qui peuvent être utilisés comme une table de correspondance pour générer la prononciation des mots sélectionnés. Dans le cas où ces ressources ne sont pas disponibles, des approches alternatives peuvent être adoptées que nous pouvons classer dans deux catégories : approches à base de règles et approches statistiques.

Les approches de la première catégorie sont basées principalement sur l'utilisation des règles linguistiques pour faire la correspondance entre les mots et la représentation phonétique. Ces approches ne capturent pas les irrégularités de prononciation dans les langues naturelles, même si des règles ou des listes d'exceptions sont incluses.

Les approches statistiques sont basées sur l'idée qu'avec suffisamment d'exemples, on peut prédire la prononciation d'un nouveau mot. Diverses techniques de modélisation ont été adaptées pour ce problème, notamment les réseaux de neurones [Yolchuyeva *et al.*, 2019, Rao *et al.*, 2015],

les arbres de décision [Udhyakumar *et al.*, 2004, Suontausta et Häkkinen, 2000], les modèles de Markov caché [Taylor, 2005a, Taylor, 2005b] et les approches statistiques pour la traduction automatique [Harrat *et al.*, 2014, Laurent *et al.*, 2009].

1.1.4 Modélisation de langage

L'un des composants les plus importants dans un système de reconnaissance automatique de la parole est le modèle de langage. Ce modèle est utilisé pour estimer la probabilité d'une séquence de mots. Cette estimation est très importante car en utilisant seulement le modèle acoustique, le signal de la parole est transcrit dans la plupart du temps en plusieurs séquences de mots ayant la même représentation phonétique. Le modèle de langage permet de choisir parmi la liste des candidats la séquence de mots la plus probable.

Un modèle de langage statistique [Manning et Schütze, 1999] est une distribution de probabilité $P(W_1^l)$ sur les séquences de mots W_1^l qui tente de refléter la fréquence à laquelle elles apparaissent comme une phrase. En utilisant la définition des probabilités conditionnelles, cette probabilité est estimée en fonction de la probabilité de chaque mot sachant l'historique de ce mot (tous les mots qui le précèdent) comme il est illustré par l'équation 1.9.

$$P(W_1^l) = P(w_1 w_2 \dots w_l) = P(w_1) \prod_{j=2}^{l} P(w_j | w_1 w_2 \dots w_{j-1}) \tag{1.9}$$

Le calcul de la probabilité conditionnelle de chaque mot w_j dépend de la fréquence d'apparition de la suite $w_1 w_2 \dots w_{j-1} w_j$ dans un corpus textuel.

$$P(w_j | w_1 w_2 \dots w_{j-1}) = \frac{N(w_1 w_2 \dots w_j)}{N(w_1 w_2 \dots w_{j-1})} \tag{1.10}$$

où $N(x)$ renvoie le nombre d'occurrences de la séquence x dans le corpus d'apprentissage. En pratique, il semble impossible de trouver un corpus contenant tous les historiques possibles pour un mot w_j, ce qui rend impossible l'estimation des probabilités $P(w_j | w_1 w_2 \dots w_{j-1})$. L'un des modèles largement utilisé pour remédier à ce problème est le modèle n-gramme.

Modèles n-grammes

Dans le modèle n-gramme, on fait l'hypothèse d'indépendance de Markov où chaque mot w_j ne dépend que des $n-1$ mots qui le précèdent $(w_{j-(n-1)} w_{j-(n-2)} \dots w_{j-1})$. Cette hypothèse permet la réduction de la taille de l'historique, ce qui permet par conséquent, d'avoir plus de chances d'observer les séquences de tailles réduites. Cette modélisation est formulée par l'équation 1.11

$$P(W_1^l) = P(w_1 w_2 \dots w_l) = P(w_1) \prod_{j=2}^{l} P(w_j | w_{j-(n-1)} w_{j-(n-2)} \dots w_{j-1}) \tag{1.11}$$

Le modèle n-gramme ne règle pas d'une manière définitive le problème d'insuffisance de données. En effet, si un mot w_j et son historique (les $n-1$ mots qui le précèdent) n'apparaissent pas dans le corpus d'apprentissage, le modèle n-gramme attribue une probabilité nulle à la séquence $w_j w_{j-(n-1)} w_{j-(n-2)} \dots w_{j-1}$, même si cette séquence est une suite de mots possible dans la langue considérée. Pour cette raison, plusieurs techniques de lissage ont été proposées qui permettent d'estimer au mieux les probabilités quand les données ne sont pas suffisantes.

L'idée des techniques de lissage est la suivante : dans le cas où un n-gramme n'est pas observé dans le corpus d'apprentissage, le modèle n-gramme se replie sur le n-gramme d'ordre inférieur $n-1$. Ce processus sera réitéré jusqu'au niveau le plus bas, c'est-à-dire le zéro-gramme où on attribue une constante indépendante du mot w_j à la probabilité conditionnelle $P(w_j|w_{j-(n-1)}w_{j-(n-2)}\dots wj-1)$. Les techniques de lissage les plus utilisées en pratique sont celles proposées par [Witten et Bell, 1991] (*Witten-Bell*) et [Chen et Goodman, 1996] (*Kneser ney*).

Les modèles de langage de type n-gramme sont considérés comme des modèles locaux, c'est-à-dire ils ne permettent pas de trouver les dépendances entre les mots éloignés. En pratique, avec la puissance de calcul offerte par les ordinateurs actuels, on peut obtenir des listes allant jusqu'à 6 mots. Cela signifie qu'un mot doit être à 6 mots de distance maximum d'un autre mot pour capturer le contexte. Ce problème se pose particulièrement dans les accords en genre et en nombre. Ainsi, les modèles n-gramme ne capturent pas facilement la relation sémantique entre les mots. Par exemple, si on suppose que la séquence de mots يمشي القط في الحديقة *ymšy ālqṭ fy ālḥdyqh* (Le chat marche dans le jardin) a été observée dans le corpus d'apprentissage tandis que la séquence مشَى الكلب في المتنزه *mšā ālklb fy ālmtnzh* (Le chien marchait dans le parc) n'a pas été observée. Le modèle n-gramme n'est pas capable de capturer la relation sémantique entre les deux séquences de mots malgré qu'à partir de la première séquence on peut facilement conclure que la deuxième séquence est très probable dans la langue arabe, car les mots يمشي *ymšy* et مشَى *mšā* (marche et marchait), القط *ālqṭ* et الكلب *ālklb* (chat et chien), et الحديقة *ālḥdyqh* et المتنزه *ālmtnzh* (jardin et parc) ont des rôles sémantiques et/ou grammaticaux similaires. Les modèles à base de réseaux de neurones sont utilisés pour résoudre ces problèmes.

Modèles basés sur les réseaux de neurones

Une première modélisation de langage à base des réseaux de neurones artificiels a été proposée par [Bengio *et al.*, 2003]. Les auteurs ont utilisé un réseau de neurones multicouches à propagation avant, *feedforward neural network*, pour l'estimation de la probabilité d'une séquence de mots. Bien que cette modélisation a montré de meilleurs résultats que les modèles n-grammes, le contexte utilisé pour le calcul des probabilités conditionnelles est toujours fixe. Le auteurs de [Mikolov *et al.*, 2010] ont proposé d'utiliser les réseaux de neurones récurrents, *Recurrent Neural Networks* (RNN), pour la modélisation de langage.

Dans les réseaux de neurones multicouches, les entrées sont supposées indépendantes et de taille fixe. Pour chaque entrée, on essaie prédire une probabilité, une classe, etc. Ces réseaux ne sont pas adaptés pour les données séquentielles où chaque entrée est de taille variable et chaque élément composant ces entrée dépend des éléments précédents. Les réseaux de neurones récurrents sont particulièrement adaptés aux applications faisant intervenir le contexte, et plus particulièrement au traitement des séquences temporelles. Cette architecture est très adaptée pour la modélisation de langage où chaque mot dépend des mots qui le précèdent. Le fonctionnement des réseaux de neurones récurrents est plus proche du vrai fonctionnement du cerveau humain qui essaie de prédire un mot dans une séquence de mots en se basant sur son contexte.

La structure générale d'un réseau de neurones récurrent est décrite par la figure 1.9

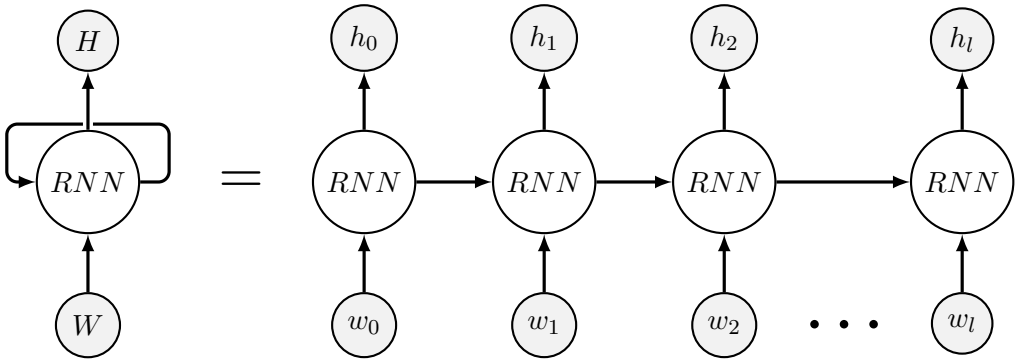


FIGURE 1.9 – Architecture d'un réseau de neurones récurrent.

Dans le cas de la modélisation de langage, chaque cellule du réseau de neurone récurrent prend en entrée un mot w_i de la séquence $W = w_1 w_2 \dots w_l$, et elle génère un vecteur h_i selon la formule 1.12.

$$h_i = \tanh\left(W_{wh} w_i + W_{hh} h_{i-1} + b_h\right) \quad \text{si } i \geq 1 \tag{1.12}$$

en pratique, le terme h_0 peut être initialisé d'une manière aléatoire ou à zéro. Le calcul de chaque vecteur h_i peut être interprété comme suit : en se basant sur ce qu'on a vu en entrée (w_i), sur la sortie de la cellule précédente h_{i-1}, et sur ce qu'on a appris lors de la phase de l'apprentissage (les matrices W_{wh}, W_{hh} et le biais b_h), on doit calculer la sortie h_i.

La sortie finale du réseau de neurones récurent est un vecteur qui associe pour chaque mot w_i une probabilité d'être le suivant dans la séquence, étant donné les $i-1$ mots précédents. Pour obtenir une telle distribution de probabilité, une couche de neurones est empilée à la suite de la couche du réseau de neurones récurent tout en utilisant comme fonction d'activation la fonction *softmax*.

1.1.5 Décodeur

En se basant sur les connaissances acquises lors de l'apprentissage de différents modèles, le décodeur génère pour un signal acoustique la séquence de mots $\hat{W}$ qui maximise la probabilité $P(W|O)$. En d'autre terme, il permet d'implanter la fonction $\arg\max$ de l'équation 1.1 et de trouver la meilleure séquence d'états génératrice. Le meilleur moyen pour trouver cette dernière est d'utiliser l'algorithme de *Viterbi* [Forney, 1973] dont le principe est le suivant : on considère la probabilité $\phi_t^{(j)} = \max_\Theta P(O_{1:t}, \theta_t = j|\lambda)$ qui représente la probabilité maximale d'observer la séquence partielle $O = o_1 o_2 \dots o_t$ et être à l'état j à l'instant t sachant les paramètres du modèle λ. Cette probabilité peut être estimée d'une manière récursive comme le montre l'équation 1.13.

$$\phi_t^{(j)} = \max_i \{\phi_{t-1}^{(i)} a_{ij}\} b_i(o-t) \tag{1.13}$$

le terme $\phi_0^{(j)}$ est initialisé à 1 pour les états initiaux non-émettrices et à 0 pour tous les autres états. En calculant les termes $\phi_t^{(j)}$ d'une manière récursive selon l'équation 1.13, La probabilité de la séquence d'états la plus probable est donnée par le terme $max_j\{\phi_T^{(j)}\}$. De plus, si chaque décision de maximisation est sauvegardée, on peut facilement trouver la séquence d'états qui maximise la probabilité.

Dans le cas de la parole continu où plusieurs modèles sont nécessaires pour le calcul du score final, l'implémentation directe de l'algorithme de Viterbi rend l'exploitation de l'espace

de recherche complexe. En pratique, les transducteurs à états finis pondérés, *weighted finite-state transducer* (WFST) [Mohri *et al.*, 2008] sont utilisés pour intégrer toutes les informations nécessaires au décodage (modèle acoustique, modèle du langage, le lexique de prononciation) dans un seul énorme graphe. Les WFSTs sont des automates dont les transitions entre les états sont étiquetées avec des symboles d'entrée et de sortie et des poids. Les poids peuvent représenter des probabilités, des pénalités ou toute autre quantité qui s'accumule le long des chemins pour calculer le score final d'association d'une chaîne de sortie à une chaîne d'entrée. Dans ce contexte, le problème de reconnaissance automatique de la parole est un problème de recherche de meilleur chemin dans un graphe qui s'appelle le graphe de décodage.

1.2 Modèles de bout en bout

Dans les approches à base du modèle de Markov caché, trois composants sont essentiels pour le développement du système de reconnaissance de la parole, à savoir : le modèle acoustique, le modèle de langage et le lexique. Ces composants sont modélisés et entraînés indépendamment, et ils sont rassemblés par le décodeur pour la génération des séquences de mots. Ces dernières années une autre approche a été proposée, qui consiste à remplacer les trois composantes de la modélisation précédente par un seul modèle basé sur les réseaux de neurones, il s'agit des modèles de bout en bout, *end2end*.

Le principe des modèles de bout en bout est d'intégrer tous les modèles dans un seul composant basé sur les réseaux de neurones récurrents. Cette nouvelle architecture tire avantage de la grande quantité de données disponibles pour rendre l'apprentissage plus efficace, en permettant une meilleure optimisation du système dans son ensemble. À titre d'exemple, [Hannun *et al.*, 2014] ont utilisé plus de 5000 heures pour entraîner leur modèle et pour dépasser les résultats obtenus avec les modèles DNN-HMM [Hinton *et al.*, 2012a].

Il existe deux approches pour développer un système de bout en bout : l'approche basée sur la classification temporelle connexionniste, *Connectionist Temporal Classification* (CTC) [Graves *et al.*, 2006] et les approches basées sur les modèles séquence-à-séquence, *sequence to sequence* (seq2seq) [Sutskever *et al.*, 2014].

1.2.1 Modèles basés sur CTC

Contrairement aux modèles DNN-HMM où l'apprentissage est basé sur des données alignées (observations acoustiques alignées avec les unités acoustiques), la classification CTC est utilisée pour entraîner des réseaux de neurones récurrents à étiqueter des séquences sans avoir besoin d'un alignement explicite entre la séquence d'entrée et celle de sortie.

Cette approche sans segmentation se rapproche des procédures d'entraînement des modèles HMM par le fait que l'alignement entre la séquence d'observation et les unités acoustiques (plus souvent des caractères dans ce cas) est basé sur des algorithmes de programmation dynamique (par exemple l'algorithme de Baum-Welch ou celui de Viterbi). L'avantage par rapport aux modèles HMM, c'est qu'en utilisant les réseaux de neurones récurrents, on peut capturer plusieurs informations à la fois, à savoir la classification acoustique et la structure linguistique, d'où le nom des modèles de bout en bout.

Pour comprendre comment la CTC est utilisée pour aligner deux séquences, considérons l'exemple suivant : on veut aligner l'observation acoustique $O = o_1 \dots o_{12}$ avec la séquence de caractères $C =$ قطة *qṭh* (chat) en supposant que le vocabulaire est composé de caractères suivants $V =$ { ق,ا,ة,ط,ب}. Pour aligner O et C, deux suppositions sont faites :

— L'observation acoustique O ne sera alignée qu'avec les caractères qui composent la séquence C. Dans notre cas, on considère seulement l'ensemble de trois caractères $V' = \{$ ق,ة,ط$\}$.

— L'alignement entre les deux séquences O et C est monotonique ; on doit ordonner les caractères composant C afin de commencer par le premier caractère (ق) et terminer par le dernier caractère (ة).

Ces deux propositions sont illustrées par la figure 1.10. À chaque instant t, le réseau de neurones prend en entrée une trame o_t et essaie de prédire le caractère qui correspond à cette trame (ق,ة ou ط). D'une manière plus formelle, il estime la probabilité de générer un caractère $c_i \in V'$ sachant l'observation O ($y_t^{c_t} = p_t(c_t|O)$).

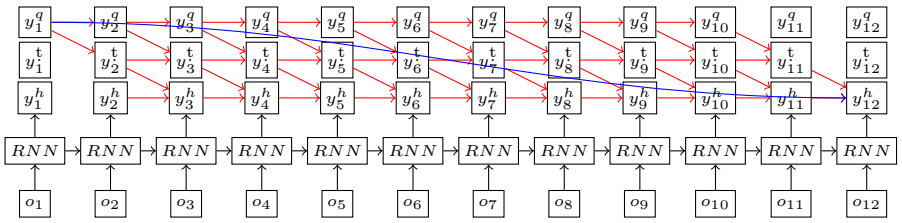


FIGURE 1.10 – Un exemple d'alignement d'une observation acoustique O avec une séquence de caractères en utilisant la CTC. L'exposant q correspond à la lettre ق, le ṭ correspond à la lettre ط et le h correspond à la lettre ة du vocabulaire V'

Après avoir considéré les deux hypothèses citées ci-dessus, un graphe est construit (les arcs en rouge dans la figure 1.10) où le score de chaque nœud est la probabilité estimée par le réseau de neurones. Pour trouver la meilleure séquence de caractères qui correspond à l'observation acoustique en entrée, un algorithme de programmation dynamique (de Viterbi par exemple) est utilisé. Cette procédure d'alignement associe pour chaque trame o_t un caractère $c \in V'$ pour obtenir, enfin, une séquence de caractères qui ressemble à ceci : $C' =$ قققققططططهههه où le nombre de caractères dans C' correspond exactement au nombre de trames dans l'observation O. Pour trouver la séquence de mots, il suffit d'éliminer les caractères en double dans C'. Cette solution pose deux problèmes : si un mot est composé de deux caractères identiques qui se suivent, par exemple le mots آآكلت *āākltt* (As-tu mangé ?), on peut jamais trouver ce mot à cause des deux caractères آآ *āā* qui seront remplacés par un seul ا *ā*. Le deuxième problème concerne le silence ou le bruit qui peuvent se produire dans la parole. Il peut y avoir des trames dans l'observation acoustique où aucun caractère n'est associe à ces trames. Pour résoudre ces deux problèmes, on étend le vocabulaire avec un caractère spécial le *blanc* ϵ. Ce caractère sera associé à chaque trame qui correspond à un silence ou pour séparer deux caractères qui se suivent pour ne pas éliminer les caractères en double.
La procédure d'alignement est utilisée pour aligner toutes les séquences dans le corpus d'apprentissage et pour pouvoir, par la suite, optimiser les poids du réseau de neurones.

Pour toute nouvelle séquence d'observation $O = o_1 \dots o_T$, l'approche pour trouver la meilleure séquence de caractères $C = c_1 \dots c_T$ qui lui correspond est basée sur un algorithme de recherche en faisceau, *Beam search*. Le score final est calculé en fonction des probabilités conditionnelles

estimées par le réseau de neurones.

$$P(W|O) = \prod_{t=1}^{T} p_t(c_t|o_t) \tag{1.14}$$

avec $c_t \in V$. En pratique, d'autres probabilités (celle du modèle de langage par exemple) ou scores (celui du "pénalité" de mots par exemple) peuvent être intégrés dans le calcul du score final pour améliorer la prédiction des séquences de mots [Hannun *et al.*, 2014, Graves et Jaitly, 2014, Amodei *et al.*, 2016, Bernath *et al.*, 2018].

1.2.2 Modèle séquence-à-séquence

Les modèles à base CTC ressemblent au modèle acoustique dans les systèmes traditionnels de reconnaissance automatique de la parole. En effet, pour chaque trame on essaie de prédire le caractère qui lui correspond. Dans les modèles séquence-à-séquence, il existe deux principaux composants : un encodeur, *encoder*, et un décodeur, *decoder*. L'encodeur prend en entrée la séquence d'observation $O = o_1 \dots o_T$ et il génère une représentation intermédiaire pour cette observation en se basant sur plusieurs blocs d'un réseau de neurones récurrent. Cette représentation intermédiaire de l'entrée, qui n'est d'autre que le vecteur produit par le dernier bloc du réseau de neurones récurrent, est utilisée par le décodeur pour prédire la séquence de caractères correspondant au mieux à l'observation O. Cette architecture est décrite par la figure 1.11.

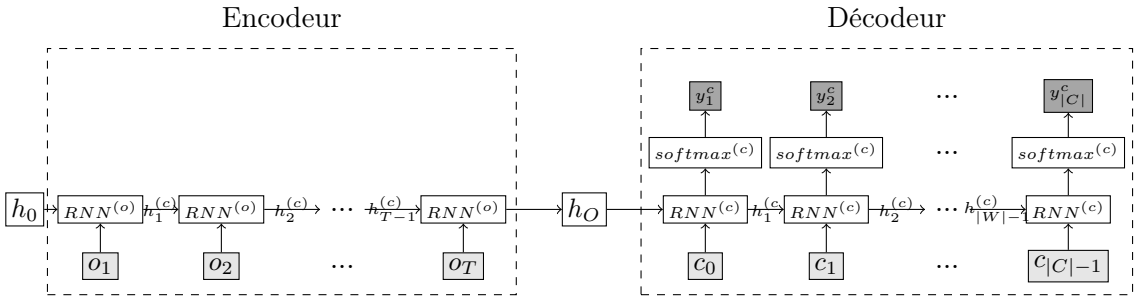


FIGURE 1.11 – Architecture du modèle séquence-à-séquence.

Cette modélisation basée sur les deux composants encodeur-décodeur a du mal à faire des généralisations pour les nouvelles séquences d'observations. Cela est justifié par le fait de se baser sur une seule représentation de la séquence d'entrée pour la génération de la séquence de sortie. Cette représentation intermédiaire est pauvre, elle ne porte aucune information explicite sur les parties les plus importantes du signal sur lesquelles on peut se baser pour mieux reconnaître la parole. Pour pallier ce problème, au lieu d'encoder la séquence d'entrée dans un seul vecteur, on utilise plusieurs vecteurs, appelés des vecteurs de contexte, pour prédire chaque caractère de la séquence de sortie ; c'est l'idée des techniques d'attention [Bahdanau *et al.*, 2014]. Chaque vecteur de contexte vc_t est calculé, comme le montre la figure 1.12 en fonction des sorties intermédiaires des blocs du réseaux de neurones récurrents de l'encodeur $h_t^{(o)},\ t \in [0, T]$ et du décodeur $h_j^{(c)},\ j \in [0, i-1]$.

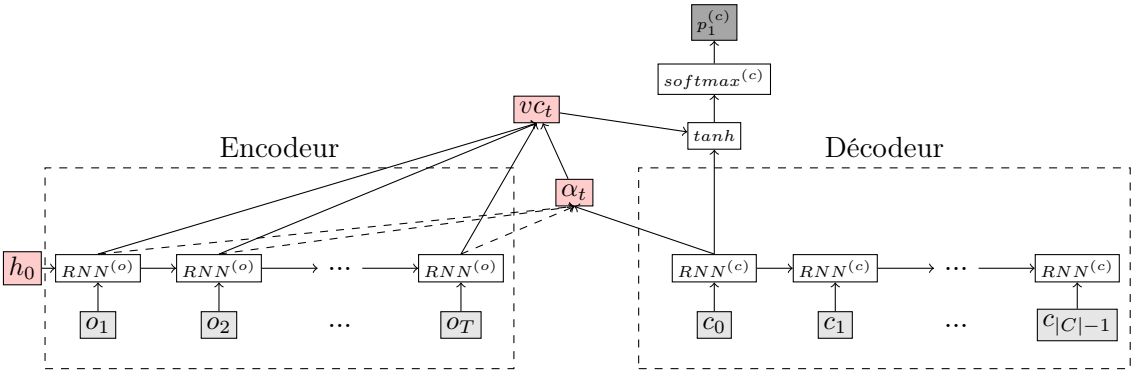


FIGURE 1.12 – Utilisation du mécanisme d'attention dans les modèles séquence-à-séquence.

Avec cette approche, l'alignement entre la séquence d'entrée et celle de sortie est modélisé d'une manière explicite. Le terme α_t dans la figure 1.12 représente la probabilité d'alignement du caractère c_t avec les observations de la séquence d'entrée. Dans ce cas, pour chaque caractère de la séquence de sortie, le réseau de neurones fait attention aux parties les plus importantes du signal qui peuvent générer cette unité, d'où le nom de la technique d'attention. Un exemple d'alignement entre le signal acoustique et la séquence de mots أَتَاحت لِلبَائِع المتجول *ā'tāḥt lilbāā'* × *ālmtǧwl* (Elle a permet au vendeur ambulant) est donné dans la figure 1.13.

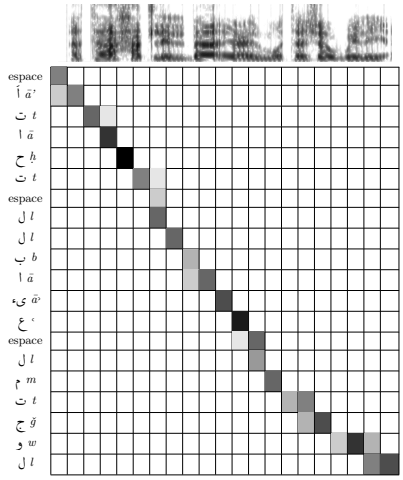


FIGURE 1.13 – Exemple d'un alignement entre le signal acoustique et la séquence de caractères généré par un modèle séquence-à-séquence.

1.3 Reconnaissance automatique de la parole arabe

Les techniques standards de modélisation acoustique et linguistique qui sont proposées pour d'autres langues, comme le français ou l'anglais, peuvent être facilement appliquées à l'arabe. En revanche, certains aspects de la langue arabe posent des problèmes que nous devons surmonter lors du développement de systèmes de reconnaissance automatique de la parole. Parmi ces aspects on cite l'absence des diacritiques (voyelles) dans le texte, la complexité de sa morphologie, la simplification de l'orthographe de certains mots (par exemple l'omission du symbole *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا) et l'existence de plusieurs variantes parlées.

Les premiers systèmes de reconnaissance automatique de la parole arabe étaient des systèmes à base de vocabulaire restreint pour accomplir des tâches simples, comme par exemple la reconnaissance de chiffres ou de mots isolés. Ces systèmes étaient principalement basés sur des règles phonologiques [Imai *et al.*, 1995], ou sur la combinaison des modèles HMM avec des algorithme de classification [Bahi et Sellami, 2003, Khasawneh *et al.*, 2004, Bourouba *et al.*, 2006].

Les avancées majeures observées dans l'état de l'art des systèmes de reconnaissance automatique de la parole arabe ont été décelées grâce à la disponibilité de ressources et le soutien de certains projets de la *Defense Advanced Research Projects Agency* (DARPA). Parmi ces projets, on peut citer le projet *Effective, Affordable, Reusable Speech-to-Text* (EARS) et son successeur *Global Autonomous Language Exploitation* (GALE) qui visaient à développer des systèmes de reconnaissance automatique de la parole arabe efficaces et à rendre les ressources plus disponibles. Parmi les systèmes efficaces développés dans le cadre du projet GALE on trouve le système de [Soltau *et al.*, 2009]. Il est basé sur grand vocabulaire comprenant 737k mots et 2,5M de variantes de prononciation. Le texte est automatiquement voyellé avec une stratégie d'initialisation uniforme (*flat-start training*). Le décodage fonctionne en deux passes : une passe pour la génération des hypothèses avec un modèle de langage n-gramme et une passe de *rescoring* avec un modèle neuronal. Le système qui en résulte permet d'obtenir un taux d'erreur inférieurs à 10%. Il est à noter que le système est entraîné avec plus de 800h de données orales.

La morphologie complexe de l'arabe entraîne un taux de mots hors vocabulaire élevé par rapport à d'autres langues tel que l'anglais. Cela a un impact direct sur le taux d'erreur dans l'évaluation finale de la sortie du système. Il existe plusieurs travaux dans la littérature qui traitent ce problème. [Afify *et al.*, 2006, Xiang *et al.*, 2006, Diehl *et al.*, 2009, Ng *et al.*, 2009] proposent d'utiliser une décomposition morphologique pour réduire la taille du vocabulaire tout en couvrant un maximum de mots. Ils ont montré que les résultats obtenus avec un lexique large peuvent être obtenus avec un lexique réduit en appliquant la décomposition morphologique.

Une autre approche intéressante consiste à utiliser des modèles de langue avec une analyse morphologique. [Choueiter *et al.*, 2006] ont utilisé un modèle du langage basé sur les morphèmes générés avec un algorithme de segmentation statistique [Lee *et al.*, 2003]. Un automate d'état fini de type accepteur a été utilisé pour définir les séquences de morphèmes acceptées. En utilisant un vocabulaire moyen (moins de 64k mots) et un modèle n-gramme basé sur les morphèmes, les auteurs ont obtenu une amélioration absolue de 2,4% par rapport au modèle conventionnel basé sur les mots. En utilisant un vocabulaire plus large (800k mots), une amélioration absolue de seulement 0,2% a été obtenue.

La langue arabe est une langue consonantique, elle ne comporte que trois voyelles, chacune d'entre elles ayant une forme longue et une forme courte. Les textes formels sont généralement écrits sans voyelles courtes, par conséquent un mot peut avoir plusieurs prononciations possibles. [Kirchhoff *et al.*, 2002] ont montré que la modélisation explicite des voyelles lors de l'apprentissage du modèle acoustique assure de meilleurs résultats. Cependant, [El-Desoky *et al.*, 2009] ont montré que le meilleur système est celui développé sans les diacritiques avec une décomposition morphologique.

La majorité des travaux proposés était focalisée sur l'arabe standard moderne, *Modern Standard Arabic* (MSA), en laissant de côté les autres variantes parlées. L'arabe standard est couramment utilisée dans les livres, les journaux, les magazines, et les médias officiels. Cependant, il n'est pas utilisé dans les conversations usuelles. En l'occurrence, on utilise plutôt d'autres variantes appelées le dialecte. Ce dernier est une forme informelle de la langue propre à chaque région du monde arabe. Le dialecte arabe est principalement fondé sur l'arabe standard en écartant plusieurs contraintes morpho-syntaxiques de la langue d'origine pour laisser place à une langue informelle plus simple d'usage. Les efforts réalisés jusqu'à présent pour développer

les systèmes de reconnaissance automatique de la parole pour les dialectes arabes concernent ceux qui partagent un nombre important de mots avec l'arabe standard, à savoir l'irakien [Afify *et al.*, 2006], l'égyptien [Ali *et al.*, 2014], le qatari [Elmahdy *et al.*, 2014] et le levantin [Elmahdy *et al.*, 2012, Soltau *et al.*, 2011].

Récemment, des compétitions connues sous le nom *Multi-Genre Broadcast Challenge* (MGB) ont boosté la recherche dans le domaine de la reconnaissance automatique de la parole pour la langue arabe ainsi que ses variantes. Il y a eu 4 éditions de MGB, pour chaque édition, de nouvelles tâches sont intégrées. Dans la dernière édition MGB-5 [Ali *et al.*, 2019], deux tâches sont considérées : la reconnaissance de la parole pour le dialecte marocain et l'identification de 17 dialectes. Les données utilisées pour la reconnaissance automatique de la parole sont une collection de 13 heures du dialecte marocain extraite à partir de 93 vidéos de YouTube. Le taux d'erreur obtenu par le meilleur système est de 59,4%. Dans le cas où ce taux est calculé en se basant sur plusieurs transcriptions de référence, les auteurs ont obtenu un taux de 37,6%. Sachant que le dialecte n'est soumet à aucune convention d'écriture, cette amélioration est principalement due à l'enrichissement en formes d'écriture issues des transcriptions de référence. Il est à noter que ces transcriptions de références ont été générées manuellement.

1.4 Conclusion et discussion

Dans ce chapitre, nous avons présenté le principe de fonctionnement des systèmes de reconnaissance automatique de la parole basés sur les approches statistiques. Trois composants sont nécessaires pour le développement de ce genre de systèmes : le modèle acoustique, le modèle de langage et le lexique. Du point de vue linguistique, le modèle acoustique permet de modéliser le système phonologique de la langue. Il fait le lien entre le signal de la parole et les unités acoustiques. Le lexique transforme les unités acoustiques en une séquence de mots en se basant et le modèle de langage assure que la séquence de mots générée respecte bien les règles linguistiques de langue. Ces trois composants sont utilisés par le décodeur afin de transformer une séquence d'observations acoustiques en une séquence de mots.

Ces dernières années, une nouvelle approche qui modélise globalement l'ensemble et permet de passer de l'acoustique au texte directement, a été proposée ; il s'agit des modèles de bout en bout. La particularité de ces modèles est qu'ils sont capables d'aligner le signal avec les caractères/mots tout en apprenant les aspects linguistiques de la langue. Leur inconvénient majeur est qu'ils sont gourmands en ressources, ils nécessitent une grande quantité de données pour atteindre des résultats état de l'art.

Le dernier point présenté dans ce chapitre était l'état de l'art des travaux déjà effectués dans le domaine de la reconnaissance automatique de la parole pour la langue arabe. Sachant que les approches statistiques soient indépendantes de la langue et elles peuvent être facilement appliquées pour la langue arabe, il est nécessaire de prendre en considération les particularité de cette langue pour booster les performances des systèmes de reconnaissance automatique de la parole arabe. Parmi les aspects qui caractérisent l'arabe, on trouve l'absence des voyelles dans le texte ce qui rend leur apprentissage plus difficile, la complexité de sa morphologie et l'existence de plusieurs variantes parlées (les dialectes).

Vue que notre travail de thèse s'inscrit dans le cadre du projet AMIS dont l'un de ces objectif est la traduction automatique de vidéos, nous présentons, dans le chapitre suivant, une vue d'ensemble sur les travaux de recherche réalisés dans le domaine de la traduction automatique.

Chapitre 2

Traduction automatique

Sommaire

La traduction automatique est le processus qui consiste à convertir automatiquement un texte donné dans une langue source vers un texte dans une langue cible, tout en préservant le sens du texte d'entrée et en produisant un texte fluide dans la langue cible. Ce processus n'est pas une simple substitution mot à mot, il est bien plus complexe. Un traducteur doit interpréter et analyser tous les éléments dans le texte source et savoir comment chaque mot peut influencer un autre pour générer enfin le texte cible.

Dans ce chapitre, nous présentons les avancées menées dans le domaine de la traduction automatique. Dans un premier temps, les deux grandes familles d'approches de la traduction automatique sont présentées, à savoir les approches dites expertes qui font appel à des connaissances linguistiques, et les approches dites empiriques qui sont basées sur une analyse empirique de données textuelles. Nous décrivons par la suite les approches statistique et neuronale utilisées dans nos travaux de thèse. À la fin de ce chapitre, nous présentons les techniques d'évaluation des traductions. Le problème de l'évaluation automatique de la traduction automatique est aussi complexe que le problème de la traduction lui-même. C'est aussi dû au fait qu'il existe en général plusieurs traductions possibles pour une phrase.

2.1 Approches de la traduction automatique

Les premières idées de la traduction automatique datent du $17^{éme}$ siècle, où Leibniz et Descartes présentaient l'idée des dictionnaires automatiques qui peuvent traduire de et vers n'importe qu'elle langue. Avant le $20^{éme}$ siècle, ce ne sont que des théories, et il faut attendre les années 40 pour voir la première proposition d'un formalisme de traduction [Weaver, 1955]. Warren Weaver voyait le problème de traduction comme un problème de décryptage d'un message crypté. Ce qui donne la possibilité d'utiliser les ordinateurs pour automatiser ce travail. Cette idée a été considérée comme responsable de la naissance de la traduction automatique comme un domaine de recherche. Quelques années plus tard, en 1954, une première expérience menée par IBM montrait la traduction d'une soixantaine de phrases russes soigneusement choisies en anglais en utilisant un vocabulaire restreint de 250 mots et une grammaire de 6 règles seulement [Hutchins, 2004a]. Bien qu'aujourd'hui un tel système semble très limité, à l'époque il a été considéré comme une réelle démonstration de la possibilité de la traduction automatique.

Après cette période, plusieurs approches de traduction automatique ont été proposées et elle peuvent être classées selon deux grandes catégories : les approches expertes et les approches empiriques.

2.1.1 Approches expertes

Les méthodes basées sur les approches expertes font appel à des connaissances linguistiques établies par des experts humains. Le processus de la traduction commence par une étape d'analyse des corpus de la langue source et ceux de la langue cible afin d'en extraire des règles linguistiques. Ces dernières représentent des règles de représentation des phrases et de transfert. Pour traduire un texte, une représentation intermédiaire de ce texte est extraite en se basant sur les règles déjà établies. Le niveau d'abstraction de cette représentation dépend de l'approche utilisée. Les principales approches qui existent sont : la traduction directe, la traduction à base de règles de transfert et la traduction par interlangue ; chacune de ces approches sera détaillée dans ce qui suit. La dernière étape dans le processus de traduction consiste à utiliser les règles de transfert pour générer à partir de la représentation intermédiaire du texte source une autre représentation équivalente dans la langue cible. Pour avoir la traduction finale, on peut faire appel à d'autres ressources linguistiques comme les dictionnaires. Pour résumer, le processus de traduction à base des approches expertes suit typiquement les trois étapes suivantes :

— Analyse : lire, analyser et comprendre le texte source pour l'exprimer en des représentations intermédiaires.
— Transfert : les représentations intermédiaires sont transférées vers d'autres représentations en langue cible.
— Génération : générer le nouveau texte en langue cible en partant des représentations intermédiaires .

Traduction directe

Dans les systèmes de traduction directe, les deux phases d'analyse et de génération sont négligées. Aucune représentation intermédiaire n'est utilisée ; on travaille directement sur le texte d'entrée et celui de sortie. Pour traduire une phrase, on procède d'abord à la segmentation lexicale de cette phrase (souvent une segmentation en mots), et à la traduction mot à mot en utilisant une table bilingue. Cette table spécifie l'ensemble des règles de traduction et de réordonnancement qui permettent de traduire les mots sources et de réordonner ceux de la phrase cible. Vu que ce

processus de traduction est très simple et intuitif, une phase de post-traitement par des experts humains est nécessaire pour corriger éventuellement les erreurs de traduction.

Ce type de système est efficace dans certains cas d'application avec un vocabulaire restreint où une traduction correcte peut être produite avec une simple substitution de mots, c'est le cas des systèmes de première génération (les systèmes de traduction russe-anglais et anglais-russe).

Traduction à base de règles de transfert

Contrairement aux systèmes de traduction directe, les systèmes à base de règles de transfert ont besoin de trois phases de traduction : l'analyse, le transfert et la génération. Le processus de traduction commence par l'analyse syntaxique, morphologique ou sémantique de la phrase source. Cette analyse conduit à une représentation intermédiaire arborescente qui va être convertie dans la langue cible lors de la phase de transfert.

Le transfert entre la langue source et la langue cible peut être effectué soit au même niveau d'abstraction ou à deux niveaux différents. Dans le premier cas, on peut passer, par exemple, de l'analyse syntaxique de la phrase source à l'analyse syntaxique de la phrase cible. Dans le deuxième cas, le transfert peut être de nature descendante, c'est-à-dire on passe d'une structure source intermédiaire à un niveau moins abstrait (par exemple : le passage de la représentation syntaxique de la phrase source à la représentation morphologique de la phrase cible). Il existe aussi le transfert de nature ascendante où on passe à un niveau plus abstrait d'une représentation cible (par exemple : le passage de la représentation syntaxique de la phrase source à la représentation sémantique de la phrase cible).

Pour générer, enfin, la traduction de la phrase source à partir de la représentation intermédiaire, une grammaire de la langue cible est utilisée. La figure 2.1 illustre un simple exemple de traduction de la phrase arabe نِهَاية سعيدة *nhāyh sʿydh*[8] vers la langue française *une fin heureuse*.

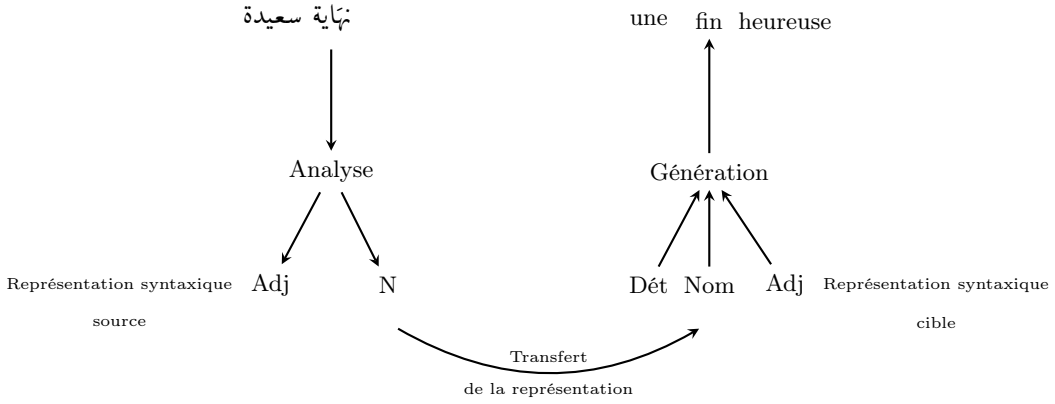


FIGURE 2.1 – Exemple de traduction avec un système à base de règles de transfert.

Dans cet exemple, la phrase source est analysée syntaxiquement pour générer la représentation intermédiaire. Selon la paire de langues utilisée dans la traduction, on peut avoir une analyse plus profonde de la phrase source.

L'avantage des systèmes de traduction automatique à base de règles de transfert est la réutilisation des règles de transfert et de la grammaire pour les nouvelles paires de langues. Ils sont également efficaces quand le vocabulaire utilisé est restreint, comme par exemple, dans le domaine

8. Il est à noter que texte arabe se lit de gauche à droite.

de la réservation d'hôtels. Cependant, pour les paires de langues où les phrases sont grammaticalement complexes, on risque de ne pas trouver la bonne représentation intermédiaire, ce qui a un impact direct sur la qualité de la traduction.

Traduction par pivot

Cette approche est fondée sur une représentation intermédiaire entre la langue source et la langue cible, connue sous le nom de représentation pivot. Dans ce cas, pour traduire d'une langue source vers une langue cible, on doit passer par cette représentation. Le problème de la traduction peut être vu, dans ce cas, comme une combinaison de deux problèmes : le premier consiste à transformer la phrase source à une représentation pivot. Par la suite, il faut générer la phrase cible à partir de la représentation pivot.

Le niveau d'abstraction de l'analyse du texte source dans la traduction automatique à base d'interlangue est le plus élevé comparé aux deux autres approches (traduction directe et traduction à base de règles de transfert). La complexité ou la difficulté de cette approche réside dans le fait de construire un vocabulaire pivot de tous les concepts possibles pour les deux langues concernées par la traduction.

Comme exemple d'une représentation pivot, on trouve *Universal Networking Language* (UNL) proposée par [Uchida et Zhu, 2001]. L'UNL se compose des mots universels, des relations, des attributs, et d'une base de connaissances. Les mots universels constituent le vocabulaire de l'UNL, les relations et les attributs constituent la syntaxe et la base de connaissances constitue la sémantique de l'UNL.

L'avantage de cette approche est le même que celui de l'approche à base de règles de transfert, à savoir la réutilisation de la langue pivot pour une nouvelle paire de langues. Ainsi, cette approche est très utile dans le cas où peu de données sont disponibles pour traduire d'une langue à une autre. On peut toujours passer par la représentation pivot pour contourner le problème du manque de données. En revanche, le processus de traduction dans cette approche est un processus tabulaire où les erreurs peuvent se propager d'une étape à une autre. En effet, le passage du texte source à la représentation pivot peut ne pas retranscrire parfaitement le sens original dans la langue source. L'utilisation de cette représentation (qu'on peut considérer comme pauvre) pour la génération du texte cible peut engendrer des pertes d'information.

2.1.2 Approches empiriques

L'inconvénient majeur des approches expertes est la mise en place du système de traduction qui est très coûteuse. D'une part, on doit toujours faire appel à des experts humains pour corriger les traductions ou pour générer les représentations intermédiaires. D'autre part, les programmes résultant de ces méthodes sont souvent spécifiques à une paire de langues. La création d'un nouveau programme est nécessaire si on veut créer un système de traduction pour une nouvelle paire de langues, bien que la réutilisation de certains modules soit possible.

Dans les approches empiriques, le processus de traduction est automatisé en apprenant des modèles à travers des analyses de corpus bilingues parallèles. Ces corpus sont des collections de phrases dans la langue source et leurs traductions dans la langue cible. Les approches empiriques sont principalement basées sur deux phases : l'apprentissage et le décodage. L'apprentissage consiste à extraire des connaissances linguistiques à partir des corpus parallèles. Ces connaissances seront utilisées, ensuite, par le décodeur. Ce dernier prédit pour une nouvelle séquence de mots, sa traduction dans la langue cible. Parmi les principales approches qui sont basées sur les méthodes empiriques, on peut citer la traduction basée sur les exemples, la traduction statistique

et la traduction neuronale. L'approche basée sur les exemples sera présentée brièvement dans ce qui suit, tandis que les deux autres approches seront présentées plus en détail dans les sections 2.2 et 2.3 respectivement.

Traduction basée sur les exemples

L'idée principale de cette approche est d'utiliser des exemples de bonnes traductions pour en produire d'autres [Nagao, 1984]. La première étape pour mettre en place cette approche est de construire une base contenant des exemples de traduction entre la langue source et la langue cible. La construction d'une telle ressource est basée sur l'analyse des corpus parallèles phrase par phrase. Pour chaque phrase, on essaie de détecter les segments parallèles en se basant sur des ressources linguistiques comme les dictionnaires.

Une fois la base d'exemples construite, elle sera utilisée par le décodeur pour générer de nouvelles traductions. Ce processus se déroule en trois étapes :

— Trouver les correspondances : Pour une phrase, on procède d'abord à la sélection des phrases d'exemples les plus proches à partir de la base construite lors de la phase de l'apprentissage. La sélection de ces phrases est basée essentiellement sur le calcul de la similarité entre la phrase source à traduire et les exemples de la base de données. Le calcul de la similarité peut être fait sur la phrase entière ou sur des segments de la phrase. Il existe d'autres méthodes qui proposent d'analyser la phrase source ainsi que les exemples de la base, pour comparer par la suite les arbres résultant de cette analyse. Dans le cas où aucune phrase exemple n'est trouvée, la phrase source est segmentée de sorte que des segments similaires existent dans la base d'exemples.
— Aligner les phrases : une fois les phrases exemples sélectionnées, on doit spécifier les parties des phrases qui peuvent être réutilisées dans la traduction de la phrase source.
— Combiner : la dernière étape consiste à assembler les traductions spécifiées dans l'étape précédente en réordonnant les segments pour créer la traduction de la phrase source entière.

La figure 2.2 représente un exemple de traduction de la phrase : *Il achète un livre sur la politique internationale* vers l'arabe, tout en spécifiant les trois étapes citées ci-dessus.

Phrase source	Il achète un livre sur la politique internationale	
Trouver les correspondances	Il achète une pomme	يشتري تفّاحة
	Je lis un livre sur la politique internationale	اقرأُ كتّابا عن السيّاسة الدولية
Aligner les phrases	Il achète une pomme	يشتري تفّاحة
	Je lis un livre sur la politique internationale	اقرأُ كتّابا عن السيّاسة الدولية
Combiner	Il achète un livre sur la politique internationale	يشتريكتّابا عن السيّاسة الدولية

FIGURE 2.2 – Exemple de traduction français-arabe fondée sur les exemples.

Cette approche produit des traductions tirées à partir de vrais exemples générés par des êtres humains, ce qui devrait garantir une meilleure qualité de traduction. En revanche, on risque de ne pas pouvoir produire une traduction s'il n'existe pas d'exemples de traduction dans la base correspondant au texte source.

2.2 Modélisation statistique de la traduction automatique

La première modélisation statistique de la traduction automatique a été proposée par [Brown *et al.*, 1993]. Cette approche est inspirée du modèle du canal bruité déjà utilisé dans les systèmes de la reconnaissance automatique de la parole. Rappelons que ce modèle est basé principalement sur deux composants, le premier permet de faire la correspondance entre les segments de la phrase source et ceux de la phrase cible, c'est ce qu'on appelle le modèle de traduction. Le deuxième composant, connu sous le nom du modèle de langage, assure que la phrase cible générée par le modèle de traduction respecte bien les règles linguistiques de la langue cible. Dans ce qui suit, les différents composants utilisés dans l'approche statistique de la traduction automatique seront présentés en détail.

2.2.1 Passer d'une langue à une autre

Le fonctionnement des systèmes de traduction automatique statistiques se rapproche de celui des systèmes de reconnaissance automatique de la parole en deux points. Le premier concerne le fait qu'il existe deux processus principaux pour la mise en place du système : l'apprentissage et le décodage (voir figure 2.3). Le deuxième point de ressemblance concerne la formulation mathématique dernière cette approche. En effet, pour une séquence de mots f dans une langue source, trouver une traduction e dans la langue cible consiste à trouver la meilleure séquence de mots $\hat{e}$ qui maximise la probabilité conditionnelle $P(e|f)$ comme le montre l'équation 2.1.

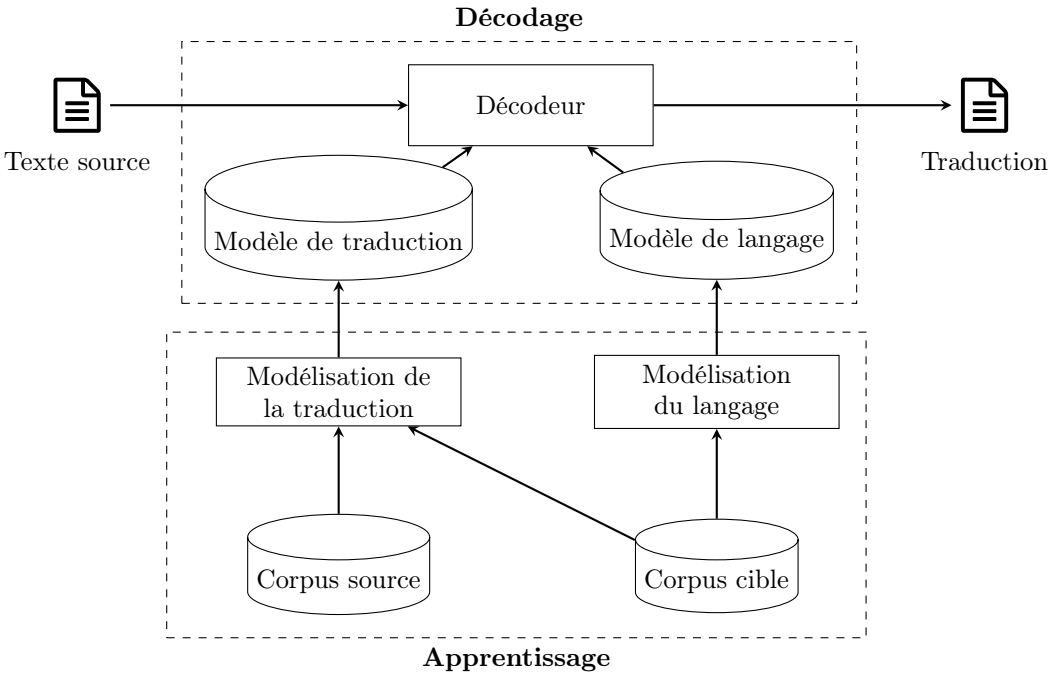


FIGURE 2.3 – Processus de la traduction automatique basé sur l'approches statistique.

$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \frac{P(f|e)P(e)}{P(f)} \tag{2.1}$$

Lors de la phase d'apprentissage, le modèle de traduction représenté par la probabilité $P(f|e)$ est entraîné sur un corpus parallèle de la langue source et de la langue cible. Le modèle de

langage qui estime la probabilité $P(e)$ est appris sur un corpus monolingue de la langue cible. L'apprentissage de ces deux modèles est basé principalement sur les fréquences d'apparition des unités lexicales dans les corpus. Le décodeur implémente la fonction arg max de l'équation 2.1 qui permet de trouver la meilleure traduction $\hat{e}$ en se basant sur les modèles appris (modèle de traduction et modèle de langage).

En pratique, les systèmes état de l'art de la traduction automatique statistique font appel à d'autres modèles en plus du modèle de traduction et celui du langage. Ces modèles sont utilisés pour capturer d'autres connaissances linguistiques qui ne sont pas modélisées par les deux modèles de base. Parmi ces modèles on peut citer à titre d'exemple le modèle de réordonnancement pour réordonner les segments dans la séquence cible, et le modèle de pénalité pour donner un avantage aux traductions longues. Dans le cas où plusieurs modèles sont utilisés, la probabilité conditionnelle finale $P(e|f)$ sera estimée en calculant le produit pondéré de tous les scores h_i générés par ces modèles comme le montre l'équation 2.2.

$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \prod_{i} h_i(f,e)^{\lambda_i} \tag{2.2}$$

avec λ_i est un coefficient de pondération associé à chaque modèle pour définir la contribution de ce dernier dans le calcul du score final. Les deux modèles les plus importants sont le modèle de traduction et celui du langage. Le principe du modèle de langage utilisé dans les systèmes de traduction automatique est le même que celui utilisé dans les systèmes de reconnaissance automatique de la parole. Il est basé sur les modèles n-grammes ou les réseaux de neurones déjà présentés dans la section 1.1.4. Dans ce qui suit, nous détaillons le modèle de traduction ainsi que le principe du décodage pour la génération de traduction.

2.2.2 Modèle de traduction

Le modèle de traduction estime la probabilité $P(f|e)$ de produire une phrase source f sachant que l'hypothèse de traduction est la phrase cible e. L'estimation de cette probabilité est basée sur le nombre de fois où la phrase f a été traduite en e dans le corpus d'apprentissage. En pratique, il est impossible de trouver un corpus parallèle qui englobe toutes les traductions possibles pour une phrase source f. La solution de ce problème consiste à apprendre des probabilités de traduction entre des segments sources et cibles et non pas entre des phrases complètes. Pour cela, [Brown *et al.*, 1993] introduit une nouvelle variable cachée a modélisant l'alignement entre les segments de la phrase source et ceux de la phrase cible. La probabilité $P(f|e)$ sera estimée, dans ce cas, en fonction de de tous les alignements possibles (A) entre les deux phrases f et e comme suit :

$$P(f|e) = \sum_{a \in A} P(f|e,a) \tag{2.3}$$

pour estimer la probabilité $P(f|e,a)$, on doit disposer de deux informations, à savoir la segmentation de la phrase source f et celle de la phrase cible e ainsi que l'alignement possible entre ces segments. Dans ce cas, la probabilité $P(f|e,a)$ est calculée comme étant le produit des probabilités de traduction des segment alignés entre l'hypothèse et la source (équation 2.4).

$$P(f|e,a) = \prod_{i=1}^{|a|} P(s_i^f | s_{a_i}^e) \tag{2.4}$$

avec s_i^f et $s_{a_i}^e$ sont respectivement le $i^{éme}$ segment de la phrase source f et son segment associé de la phrase cible avec la fonction d'alignement a. La figure 2.4 présente un exemple d'alignement entre la phrase source arabe المنزل الآخضر *ālmnzl ālā_ḥḍr* et sa traduction en français *La maison verte*.

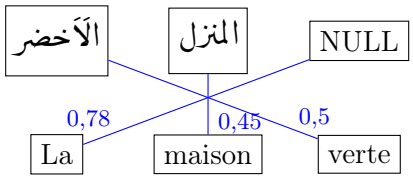


FIGURE 2.4 – Exemples de segmentations et d'alignements au niveau du mot.

Dans l'exemple de la figure 2.4, les phrases source f =المنزل الآخضر *ālmnzl ālā_ḥḍr* et cible $e = La\ maison\ verte$ sont segmentées en trois segments chacune ($s_1^f = NULL$, s_2^f =المنزل et s_3^f =الآخضر, et $s_1^e = La$, $s_2^e = maison$ et $s_3^e = verte$) et la fonction d'alignement est définie comme suit $a = \{1, 2, 3\}$. Cela signifie que le premier segment de la phrase f est aligné avec le premier de la phrase e, le deuxième avec le deuxième et ainsi de suite. La segmentation de la phrase source f est étendue par le mot vide $NULL$ pour respecter les contraintes imposées par la fonction d'alignement a qu'on expliquera par la suite. La question qui se pose maintenant, comment on peut définir la fonction d'alignement a ainsi que les différentes probabilités $P(s_i^f|s_{a_i}^e)$ d'aligner chaque segment source s_i^f avec un segment cible $s_{a_i}^e$? Bien qu'un être humain puisse facilement trouver l'alignement entre les segments de la phrase source et ceux de la phrase cible, automatiser ce processus est une tâche délicate surtout pour les paires de langue où la structure linguistique est différente. Pour y remédier, on commence par aligner les mots de la phrases source et ceux de sa traduction. Ces alignements sont utilisés, par la suite, pour trouver l'alignement au niveau de segments. Ces deux procédures sont présentées dans ce qui suit.

Trouver l'alignement au niveau des mots

La traduction d'une phrase source dans les premiers modèles proposés par [Brown *et al.*, 1993] était générée mot à mot. À partir des corpus parallèles où on dispose des alignements au niveau de la phrase, l'objectif était de trouver les alignements possibles entre les mots de la phrase source et ceux de sa traduction. Pour répondre à cet objectif, [Brown *et al.*, 1990] ont modélisé l'alignement par une variable cachée qui doit vérifier les conditions suivantes :

- Chaque mot f_j de la phrase source est connecté à exactement un et un seul mot e_i de la phrase cible.
- Plusieurs mots f_j de la phrase source peuvent être liés au même mot e_i de la phrase cible.
- Lorsqu'un ou plusieurs mots d'une phrase n'ont pas de correspondance dans l'autre phrase, on peut utiliser un mot spécial $NULL$.

Ces contraintes sont illustrées dans la figure 2.5.

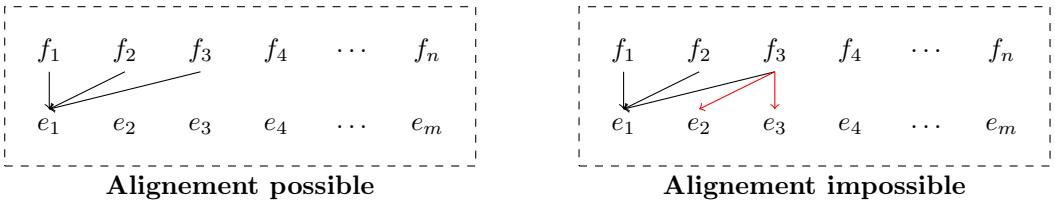


FIGURE 2.5 – Les contraintes de la fonction d'alignement au niveau de mots.

Pour trouver cette variable cachée qui modélise l'alignement et pour estimer, par conséquence, la probabilité $P(f|e)$, [Brown *et al.*, 1993] ont proposé cinq modèles connus sous le nom des modèles d'IBM. Le paramètres de ces modèles sont estimés dans un processus itératif où chaque modèle s'appuie sur les résultats du modèle qui le précède. Le premier modèle IBM-1 est le plus intuitif, il estime la probabilité lexicale d'associer chaque mot dans la phrase source avec un mot dans la phrase cible en se basant sur l'algorithme EM. Cette estimation de la probabilité lexicale est améliorée en passant aux modèles d'ordre supérieur en introduisant d'autres paramètres (réordonnancement, fertilité, insertion, etc.). Une fois que les paramètres des cinq modèles sont estimés, une table dite table de traduction est générée. Elle contient une liste de mots dans la langue source et leurs traductions dans la langue cible ainsi que les probabilités de traduction.

L'inconvénient majeur des modèles à base de mots est le fait de ne pas capturer le contexte, chaque mot est traduit indépendamment ce qui dégrade la qualité de la traduction. Un bon exemple d'une telle situation est la traduction des expressions idiomatiques. Ainsi, ces modèles sont très complexes et difficiles à implémenter ; en apprenant un modèle à base de séquences de mots, plusieurs paramètres peuvent être estimés d'une manière implicite, à savoir la notion de fertilité, de l'insertion et de la suppression qui sont définies d'une manière explicite dans les modèles d'IBM. Cette modélisation à base de séquences de mots rend le modèle beaucoup plus simple à implémenter et à utiliser.

Trouver l'alignement au niveau de segments

Une première modélisation à base de groupes de mots a vu le jour au début des années 2000 [Och, 1999, Koehn *et al.*, 2003]. Dans ces modèles, chaque groupe de mots dans la phrase source peut être aligné avec un autre groupe de mots dans la phrase cible. Cet alignement est basé sur deux alignements bidirectionnels au niveau de mots. En effet, la première étape consiste à aligner les mots de la phrase source avec ceux de la phrase cible (alignement source-cible) et inversement (alignement cible-source) comme le montre la figure 2.6. Dans cet exemple, l'alignement est représenté par une matrice de taille $n \times m$ avec n est le nombre de mots dans la phrase source et m est le nombre de mots dans la phrase cible. Chaque élément a_{ij} de la matrice indique si le $i^{éme}$ mot de la phrase source est aligné avec le $j^{éme}$ mot de la phrase cible.

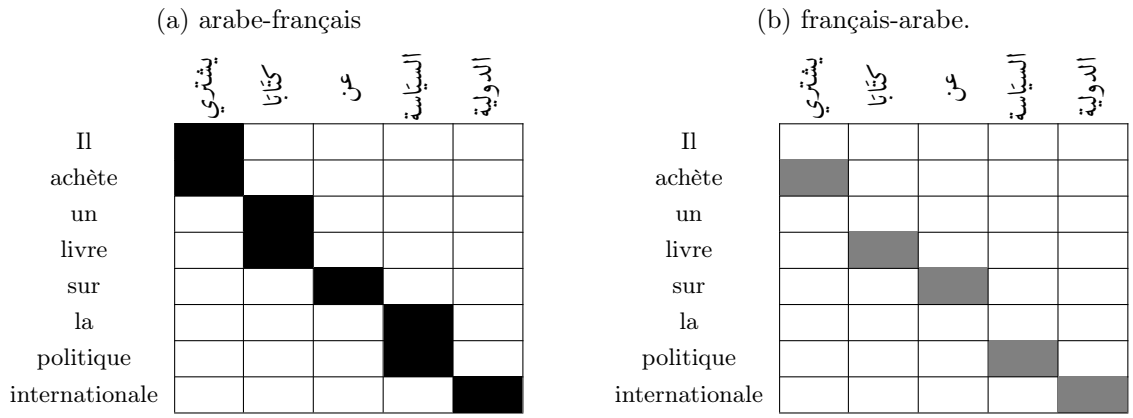


FIGURE 2.6 – Exemples d'alignement bidirectionnel entre une phrase en arabe et sa traduction en français.

L'étape suivante consiste à symétriser les deux alignements en faisant l'union et l'intersection entre les deux alignements bidirectionnels de l'étape précédente. L'intersection entre les deux alignements permet de fonder un nouvel alignement qui donne une précision élevée et un rappel faible par rapport à chaque alignement, tandis que l'union donne un alignement avec une précision faible et un rappel plus élevé [Knight et Koehn, 2003]. Un exemple d'alignement symétrisé est présenté par la figure 2.7

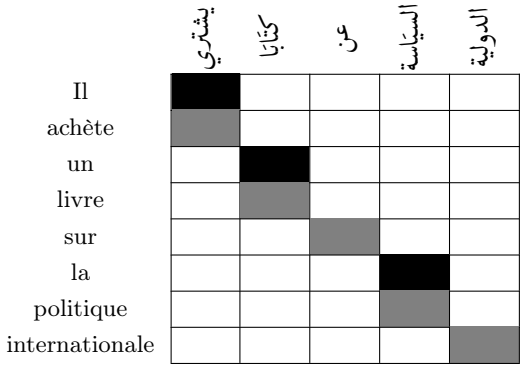


FIGURE 2.7 – Exemple d'un alignement symétrique (union/intersection).

Une fois la matrice qui représente l'alignement symétrisé faite, on peut procéder à l'extraction des groupes de mots. Cette étape consiste à extraire les correspondances en groupes de mots entre la phrase source et sa traduction à partir de l'alignement symétrisé. Ces correspondances doivent être consistantes avec l'alignement symétrisé, c'est-à-dire les mots dans une correspondance sont alignés les uns aux autres, et non pas à des mots de l'extérieur. D'une manière plus formelle :

$$
\begin{aligned}
(\tilde{e}, \tilde{f})\ \textit{consistant avec l'alignement}\ A \quad & \Leftrightarrow \\
& \forall e_i \in \tilde{e} : (e_i, f_j) \in A \Rightarrow f_j \in \tilde{f} \\
\textit{et} \quad & \forall f_j \in \tilde{f} : (e_i, f_j) \in A \Rightarrow e_i \in \tilde{e} \\
\textit{et} \quad & \exists e_i \in \tilde{e},\ f_j \in \tilde{f} : (e_i, f_j) \in A
\end{aligned}
\tag{2.5}
$$

$\tilde{f}$ et $\tilde{e}$ sont des segments de la phrase source et de la phrase cible respectivement. En respectant ces contraintes, on regroupe les cases adjacentes cochées dans la matrice qui représente l'aligne-

ment symétrisé afin de produire des segments et des alignements entre eux. On regroupe aussi les segments adjacents dans la matrice pour produire des segments de tailles plus grandes. En pratique la taille maximale des segments est souvent fixée à 7 mots par segment [Koehn *et al.*, 2003]. Tout comme les modèles à base de mots, les modèles à base de segments reposent sur une table de traduction dans laquelle des segments de la langue source et leurs traductions sont sauvegardés. Pour chaque traduction, plusieurs scores sont aussi sauvegardés(le score de traduction de $\tilde{e}$ en $\tilde{f}$, le score de $\tilde{e}$ pour le modèle de langage, le score de réordonnancement, etc.) pour faciliter le calcul du score final lors du processus du décodage.

2.2.3 Décodeur

Le décodeur est la partie du système qui permet de générer pour une phrase dans une langue source sa traduction dans une autre langue en se basant sur les différents modèles déjà appris. En d'autres termes, il permet d'implémenter la fonction arg max de l'équation 2.1. Plusieurs algorithmes ont été proposés dans la littérature pour résoudre ce problème [Koehn, 2004, Douib *et al.*, 2016, Langlais *et al.*, 2007], le plus utilisé est celui proposé par [Koehn *et al.*, 2007b]. Ce décodeur est basé sur un algorithme de recherche en faisceau dont l'idée est de construire la traduction d'une phrase source f de manière progressive.

Le décodeur extrait, dans un premier temps, pour la phrase source f, tous les segments possibles à partir de la table de traduction, c'est ce qu'on appelle les options de traduction. Pour chaque option, on étend, par la suite, l'espace de recherche en traduisant d'autres segments de la phrase source qui n'ont pas été traduits. On procède de la même manière pour construire l'espace de recherche tout en générant des hypothèses partielles de traduction. Ce processus est répété jusqu'à ce que tous les segments de la phrase source sont traduits.

Cette procédure explose l'espace de recherche, pour une phrase source de n mots et un vocabulaire V_e de la langue cible, le nombre maximum d'hypothèses est estimé à $2^n \times |V_e| \times n$ [Koehn *et al.*, 2003]. Cela rend le problème de traduction un problème NP-complet bien plus difficile à résoudre que celui de la reconnaissance automatique de la parole [Knight, 1999], d'où la nécessité d'utiliser des algorithmes gloutons ou de programmation dynamique. L'algorithme de recherche en faisceau [Koehn, 2004] est employé pour ne pas parcourir tout l'espace de recherche. En pratique d'autres mécanismes sont employés pour réduire encore l'espace de recherche, à savoir la fusion des hypothèses et l'élagage. Le premier mécanisme, celui de la fusion des hypothèses, est un moyen sans risque pour réduire l'espace de recherche. Il consiste à choisir le chemin le moins coûteux dans le cas où plusieurs chemins mènent à la même hypothèse. En ce qui concerne l'élagage, c'est un mécanisme qui risque d'éliminer des hypothèses potentielles. Le principe est basé sur l'élimination des hypothèses qui ont un score inférieur à un certain seuil ou de garder seulement les n-meilleures hypothèses à chaque niveau. On peut mieux estimer ce paramètre pour améliorer la qualité de la traduction.

2.3 Modélisation basée sur les réseaux de neurone

Le premier travail qui a essayé de modéliser le problème de la traduction avec les réseaux de neurones est celui proposé par [Cho *et al.*, 2014b] en 2014. Le modèle était basé sur une architecture séquence-à-séquence déjà présentée dans la section 1.2.2. La différence entre la reconnaissance automatique de la parole et la traduction automatique réside dans la représentation des entrées et des sorties. Le modèle adapté pour la traduction automatique prend en entrée la phrase source. Chaque mot de cette phrase est projeté dans espace vectoriel afin d'avoir une représentation numérique. Cette dernière sera l'entrée de l'encodeur. Le travail de [Cho *et al.*, 2014b]

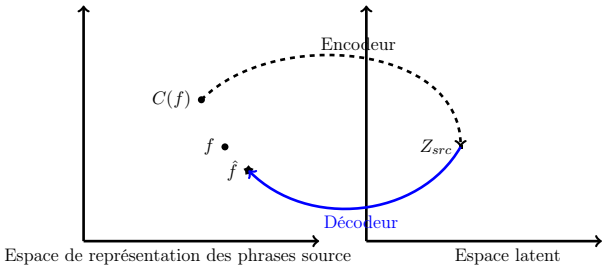
n'était pas concurrentiel par rapport à l'approche statistique à cause de la pauvre modélisation des alignements entre les mots de la phrase source et ceux de la phrase cible. En effet, la phrase cible est générée en se basant seulement sur une représentation fixe de la phrase source. Cette représentation ne prend pas en considération les parties de la phrase source les plus importantes sur lesquelles on doit se baser pour la production des mots de la phrase cible. Ce problème est résolu avec les modèles d'attention proposés par [Bahdanau *et al.*, 2014].

Au cours des années suivantes, de nombreuses avancées ont été réalisées dans ce domaine. Google [Wu *et al.*, 2016] propose son propre modèle basé sur l'architecture séquence-à-séquence. Ils utilisent huit couches de réseaux de neurones récurrents pour encoder la phrase source d'une part, et de la décoder d'autre part. Le modèle est basé sur le modèle d'attention pour la génération des mots de la phrase cible. L'autre caractéristique de ce modèle est la traduction des mots inconnus. En effet, le modèle est entraîné sur un vocabulaire restreint, tous les mots qui n'apparaissent pas dans le vocabulaire sont remplacés par un mot spécial *<unk>*. Pour y remédier, les auteurs ont proposé de décomposer le mot inconnu en des unités plus petites (des caractères par exemple) et la traduction dans ce cas sera exécutée au niveau de caractères. Ce modèle est utilisé dans leur plate-forme de traduction en ligne.

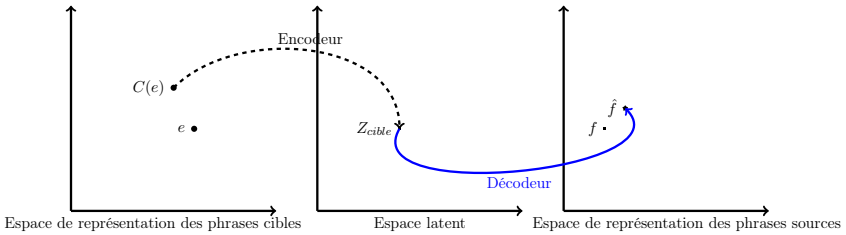
Pour certaines paires de langues, la traduction automatique basée sur les réseaux de neurones a atteint des performances impressionnantes ; il y a des cas où il est presque impossible de distinguer les traductions générées par l'approche neuronale et celles générées par un être humain [Wu *et al.*, 2016]. Cependant, ces approches nécessitent une quantité massive de données parallèles, de l'ordre de millions de phrases parallèles pour l'apprentissage des modèles. Or, les corpus parallèles sont coûteux à construire car ils nécessitent une expertise humaine, et ils sont souvent non disponibles pour les langues peu dotées en ressources. En partant de l'idée que les données monolingues sont beaucoup plus faciles à trouver, les auteurs de [Lample *et al.*, 2017] ont proposé un système de traduction où l'apprentissage est basé uniquement sur les données monolingues de la paire de langues.

L'idée de la traduction automatique basée sur les corpus monolingues est de construire un espace latent entre les deux langues et d'apprendre à reconstruire les phrases à partir de cet espace. Cette représentation latente permet de capturer la structure linguistique de chaque langue en se basant sur une architecture de type encodeur/décodeur. Le modèle proposé est entraîné sur les deux tâches suivantes [Lample *et al.*, 2017] (voir la figure 2.8) :

— Apprendre à reconstruire une phrase dans une langue donnée (source ou cible) à partir d'une version bruitée de celle-ci, comme dans les auto-encodeurs [Vincent *et al.*, 2008].
— Apprendre à reconstruire les phrases cibles à partir de la représentation latente dans le domaine source de la version bruitée de la même phrase cible, et vice versa.



(a) Reconstruction de la phrase source avec le principe de l'auto-encodeur.



(b) Passage d'une phrase cible à une phrase source (la traduction).

FIGURE 2.8 – Principe de la traduction automatique basée sur des corpus monolingues.

La figure 2.8a illustre le principe de l'auto-encodeur. Pour apprendre ce modèle, un bruit est appliqué à la phrase source f avec la fonction $C(f)$. Cette fonction applique deux transformations sur une phrase donnée : elle supprime aléatoirement des mots avec une certaine probabilité et elle applique une permutation choisie au hasard. En projetant la représentation bruitée de la phrase source f dans l'espace latent en utilisant l'encodeur, on essaie de reconstruire la phrase source en minimisant la distance entre la phrase f et la phrase reconstruite $\hat{f}$ avec le décodeur. La figure 2.8b montre le principe de la traduction. La représentation bruitée de la phrase cible $C(e)$ est projetée dans l'espace latent du domaine de la langue source. À partir de cette représentation latente, le modèle essaie de prédire la traduction $\hat{f}$ qui est proche de la phrase source f.

En apprenant à reconstruire des phrases dans les deux langues à partir d'un espace latent partagé, le modèle apprend à traduire sans utiliser de données parallèles. Les score BLEU (voir la section 2.4.2 pour plus de détails sur ce score) obtenu sur la paire de langue français-anglais est de 32,8% qui est très acceptable.

Cette approche est intéressante dans le cas de la traduction des dialectes arabes qui sont des langues peu dotées en ressources. Ils sont principalement parlés mais avec l'émergence des sites communautaires, les données monolingues des dialectes sont facilement accessibles.

2.4 Évaluation de la traduction automatique

Le problème de l'évaluation de la traduction automatique est aussi complexe que le problème de la traduction lui-même. Ceci est dû à la complexité de la langue naturelle et les ambiguïtés qu'elle présente ; une seule phrase source peut avoir plusieurs traductions totalement différentes et qui sont toutes correctes. Le besoin de l'évaluation de la traduction automatique est apparu suite au grand nombre d'approches de traduction proposées dans la littérature. Évaluer la traduction permet, d'une part, d'améliorer la sortie du système et de pouvoir comparer, d'autre part, les différentes approches.

Deux principaux critères sont utilisés pour juger la qualité de traduction : l'intelligibilité,

Fluency, et l'adéquation, *Adequacy*. Le critère d'intelligibilité est utilisé pour assurer que la sortie du système est bien écrite dans la langue cible. Tandis que, le deuxième critère celui de l'adéquation assure que l'information portée par la phrase source a été totalement transmise à la phrase cible. En se basant sur ces deux critères, un score est attribué à chaque traduction que se soit par un expert humain ou par des mesures automatiques. Le score global est calculé pour l'ensemble des traductions, afin d'estimer la qualité des traductions au niveau du corpus.

L'évaluation de la qualité de la traduction automatique dans les premiers systèmes était faite par des êtres humains qui sont souvent appelés des juges. Cependant, l'évaluation humaine ou subjective est très coûteuse en termes du temps et de ressource. Ajoutons à cela la grande quantité des traductions et le nombre élevé des systèmes à évaluer rendant le processus de l'évaluation humaine impossible. Des méthodes d'évaluation automatique ont été développées pour automatiser ce processus et remédier aux problèmes que l'évaluation humaine pose. Dans ce qui suit, nous présentons l'idée dernière l'évaluation humaine de la traduction automatique ainsi que les métriques utilisées pour automatiser ce processus.

2.4.1 Évaluation manuelle de la traduction automatique

L'évaluation manuelle, humaine , également qualifiée d'évaluation subjective, de la traduction automatique est la première méthode utilisée pour évaluer les traductions. Elle fait appel à des experts humains souvent bilingues pour juger si la traduction est correcte ou non. Dans le cas où les experts sont monolingues, on leur donne une traduction de référence générée par un expert bilingue. Leur tâche, dans ce cas, sera de mesurer la similitude entre la sortie du système et la traduction de référence.

Pour mesurer l'intelligibilité et l'adéquation, les experts doivent donner leurs avis en attribuant à chaque traduction un score allant de 1 à 5. Pour le critère d'intelligibilité, une score de 5 veut dire que la traduction est comparable à une phrase écrite par un natif de la langue cible, et un score de 1 veut dire que la traduction est totalement incompréhensible. De même pour l'adéquation, un score de 5 veut dire que la totalité du sens est conservée dans la traduction alors qu'un score de 1 veut dire qu'on ne retrouve rien du sens de la phrase source dans sa traduction.

Bien que l'évaluation humaine de la traduction automatique permette souvent d'avoir une meilleure estimation de la qualité des traductions, elle reste lourde et généralement très coûteuse en termes de temps et de coût financier. Dans certains cas, chaque traduction doit être évaluée par plusieurs juges pour minimiser le degré de subjectivité et s'assurer de la fidélité de l'évaluation. Depuis les années 2000, beaucoup de chercheurs ont proposé des méthodes automatiques pour l'évaluation de la traduction automatique.

2.4.2 Évaluation automatique de traduction automatique

L'évaluation automatique ou objective permet de traiter de grandes quantités de données, de comparer entre des instances d'un même système ou de classer les différents systèmes en termes de qualité de traduction en un temps très limité par rapport à l'évaluation humaine.

Une bonne métrique d'évaluation automatique doit être corrélée avec la décision humaine, qui reste le moyen le plus fiable pour juger la qualité de la traduction, et ceci sans l'intervention des juges humains. Ces métriques sont basées sur le calcul de la similarité entre l'hypothèse de traduction produite le système de traduction et une traduction de référence faite généralement par un être humain.

Les métriques d'évaluation automatiques de la traduction automatique sont basées principalement sur le rappel et la précision. Ces dernières sont des métriques très utilisés dans le domaine

du traitement automatique du langage naturel (TALN). Leur importance ainsi que la manière pour les calculer changent en fonction du sous domaine du TALN dans lequel elles sont utilisées. Dans la traduction automatique, les deux métriques sont équitablement importantes, car on doit assurer que le système ne produise pas des mots erronés comme il ne doit pas ignorer la traduction de certains mots [Koehn, 2009].

Le rappel représente le pourcentage de mots communs entre la traduction candidate et la traduction de référence par rapport à la taille de la traduction de référence. La précision, quant à elle, représente le pourcentage des mots en communs entre la traduction candidate et la traduction de référence par rapport à la taille de la traduction candidate.

Une bonne traduction aura une bonne précision et un bon rappel. Il existe aussi une autre métrique la *F-mesure* qui représente la moyenne harmonique du rappel et de la précision. C'est cette mesure qu'on veut maximiser pour avoir une bonne qualité de traduction.

Plusieurs autres métriques ont été proposées et sont basées sur l'optimisation du rappel et/ou de la précision. Parmi ces métriques, on peut citer : *Word Error Rate* (WER), *Position-independent word Error Rate* (PER) [Tillmann *et al.*, 1997], *Translation Error Rate* (TER) [Snover *et al.*, 200 *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) [Banerjee et Lavie, 2005] et *BiLingual Evaluation Understudy* (BLEU) [Papineni *et al.*, 2002]. Cette dernière métrique est la plus utilisée pour évaluer les systèmes de traduction automatique.

Word Error Rate (WER)

Le WER, utilisé généralement pour évaluer les systèmes de reconnaissance automatique de la parole, est l'une des premières mesures proposées pour l'évaluation de la traduction automatique. Elle se base sur la distance de Levenstein pour calculer le nombre minimum d'insertions, de substitutions et de suppressions pour rendre la traduction candidate identique à la traduction de référence. Pour avoir le score final de la mesure, le nombre minimum d'opérations à effectuer est divisé par la taille, en mots, de la traduction de référence.

$$WER = \frac{insertion + substitutions + suppressions}{|référence|} \tag{2.6}$$

La valeur du WER n'est pas bornée, le taux maximum peut dépasser 1 dans le cas il y a beaucoup d'insertions. Plus la valeur est petite meilleure est l'hypothèse de traduction. Le WER est sensible à l'ordre d'apparition des mots dans la traduction de référence et dans celle produite par le système. Supposons le scénario suivant : on dispose d'une traduction de référence R=*le chat et le chien sont des animaux domestiques.* et de trois traductions T_1, T_2 et T_3 générées respectivement par trois systèmes différents S_1, S_2 et S_3 :

— T_1 : le chat et le chien sont des animaux domestiques.
— T_2 : le chien et le chat sont des animaux domestiques.
— T_3 : domestiques des animaux sont le chat et le chien.

En calculant le WER pour chaque phrase T_i, on trouve que $WER(T_1) = 0 < WER(T_2) = 0,22 < WER(T_3) = 1$. Cela signifie que le système S_1 est meilleur que le système S_2 qui est, à son tour, meilleur que S_3 malgré que les systèmes S_1 et S_2 aient, tous les deux, produits des traductions parfaites. Une solution à ce problème consiste à comparer la traduction candidate avec plusieurs traductions de référence si elles existent, c'est ce qu'on appelle le *multiple-reference WER* (mWER). Dans le cas où on dispose d'une seule traduction de référence, la mesure PER peut être utilisée.

Position-independent word Error Rate (PER)

Le PER permet de comparer la traduction candidate et la traduction de référence en calculant le pourcentage de mots en commun entre les deux traductions sans tenir compte de l'ordre de mots [Tillmann *et al.*, 1997].

$$PER = 1 - \frac{commun - \max(0, |candidate| - |référence|)}{|référence|} \tag{2.7}$$

Avec *commun* représente le nombre de mots en commun entre la traduction de référence et la traduction candidate. Cette mesure est seulement basée sur le calcul de mots en commun, elle aura du mal à évaluer deux traductions de même longueur et composées de mêmes mots (la phrase T_3 de l'exemple ci-dessus aura un PER égale à 0 malgré que la traduction ne soit pas bonne).

Translation Edit Rate (TER)

Comme le WER, le TER est une métrique basée sur le calcul du nombre de modifications à appliquer sur la traduction candidate afin d'atteindre la traduction de référence. Cette métrique ajoute une opération en plus par rapport aux opérations considérées par le WER, il s'agit du déplacement d'un groupe contigus de mots à la fois. Chaque déplacement est compté comme une seule opération.

$$TER = \frac{insertion + substitutions + suppressions + déplacements}{|référence|} \tag{2.8}$$

En appliquant la formule 2.8 sur notre exemple, on obtient $TER(T_1) = 0$, $TER(T_2) = 0,22$ (*deux substitutions*) et $TER(T_3) = 0,33$ (*trois déplacements*). Il est clair que le TER est sensible à l'ordre de mots dans la traduction candidate.

Plusieurs autres variantes de TER existent. Parmi ces variantes on peut mentionner HTER (Human TER) [Snover *et al.*, 2006] qui est TER avec l'intervention humaine et TERp (TER plus) [Snover *et al.*, 2009] qui est TER avec l'ajout d'autres opérations qui permettent de transformer la traduction candidate en traduction de référence.

BiLingual Evaluation Understudy (BLEU)

BLEU, proposé par [Papineni *et al.*, 2002], est la métrique la plus utilisée dans le domaine de la traduction automatique pour l'évaluation des sorties des systèmes. Elle est devenue un standard adopté dans la plupart des campagnes d'évaluation. Son fonctionnement est similaire à celui de la mesure WER, mais le BLEU considère, en plus de la ressemblance mot à mot, la ressemblance au niveau des n-grammes communs entre la traduction candidate et la traduction de référence.

Le score BLEU est calculé comme suit :

$$BLEU = BP \times \exp \sum_{n=1}^{N} w_n \log p_n \tag{2.9}$$

$$BP = \begin{cases} 1 & \text{si } c > r, \\ e^{1-\frac{r}{c}} & \text{sinon.} \end{cases} \tag{2.10}$$

avec p_n est le nombre des n-grammes en commun entre la traduction candidate et la ou les références divisées par le nombre total des n-grammes de la traduction candidate. Les w_n sont des

poids associés à chaque ensemble de segments de taille n. Ces poids sont utilisés pour donner de l'importance à la probabilité p_n de certains n-grammes qui sont plus importants que d'autres. Le dernier terme *BP Brevity Penalty* est utilisé pour pénaliser les traductions courtes par rapport à la référence. En effet, dans le cas où une traduction candidate (c) est plus courte que la traduction de référence (r), le score *BLEU* sera ajusté proportionnellement à la différence entre la taille de la référence et celle de la traduction candidate.

BLEU est une mesure de précision. Elle varie entre 0 et 1, avec une valeur de 1 est attribuée à la traduction candidate dans le cas où elle est identique à la référence. La valeur de 0 est attribuée quand les deux traductions n'ont aucun n-gramme en commun. BLEU a prouvé, dans plusieurs campagnes d'évaluation, qu'elle corrèle bien avec le jugement des experts humains [Coughlin, 2003]. En revanche, il arrive des cas où la corrélation entre la métrique BLEU et le jugement humain est faible, notamment si la traduction est faite dans des domaines larges où on dispose d'une grande diversité de sujets et de phrases. BLEU a besoin, dans ce cas, de plus d'une traduction de référence ou d'un ensemble de tests volumineux afin de produire des résultats en corrélation fiable avec les évaluations humaines

Metric for Evaluation of Translation with Explicit ORdering (METEOR)

METEOR a été proposé par [Banerjee et Lavie, 2005] pour adresser quelques problèmes que le score BLEU pose. En effet, nous avons présenté au début de cette section le rôle du rappel et de la précision dans l'évaluation automatique de la traduction automatique. Les recherches ont montré que le rappel est plus important que la précision pour assurer que tout le sens a été transmis vers le résultat de la traduction. En revanche, en regardant la formule du BLEU, on peut remarquer qu'il est basé uniquement sur la précision. Ajoutons à cela le fait que BLEU ne traite les synonymes et les paraphrases que si ces derniers figurent dans l'ensemble des traductions de référence multiples [Callison-Burch *et al.*, 2006]. Dans le cas où on dispose d'une seule traduction de référence, BLEU n'est pas capable de capter le sens de mots.

En plus de la précision, [Banerjee et Lavie, 2005] ont proposé d'intégrer le rappel dans le calcul du score final pour assurer une forte corrélation avec le jugement humain. En outre, au lieu de se baser seulement sur la similarité lexicale entre les n-grammes de la traduction candidate et de la référence, METEOR utilise des dictionnaires pour capturer les synonymes des mots. Mieux encore, cette métrique peut se replier à un niveau d'analyse plus bas pour chercher des racines de mots et des classes de mots. Le problème de cette métrique est qu'elle est plus compliquée à calculer avec beaucoup plus de paramètres à estimer par rapport à BLEU.

2.5 Traduction de la langue arabe

La plupart des travaux réalisés dans le domaine de la traduction automatique de la langue arabe se sont focalisés sur la traduction de cette langue vers la langue anglaise. Peu de travaux ont été réalisés dans l'autre sens ou de n'importe quelle langue vers l'arabe. Peu importe le sens de traduction, l'arabe a un ordre de mots différent qui pose un défi important à la traduction automatique. Cela est principalement dû au nombre important de possibilités d'exprimer la même phrase en arabe. Trois éléments principaux composent les phrases arabes, à savoir le sujet, le verbe et l'objet. Selon l'ordre de ces éléments dans la phrase, cette dernière peut être classée en quatre catégories : SVO, VSO, VOS et SOV [Alqudsi *et al.*, 2012]. Cela rend le réordonnancement entre la phrase arabe et la phrase de l'autre langue plus complexe pour les systèmes de traduction automatique (voir la figure 2.9).

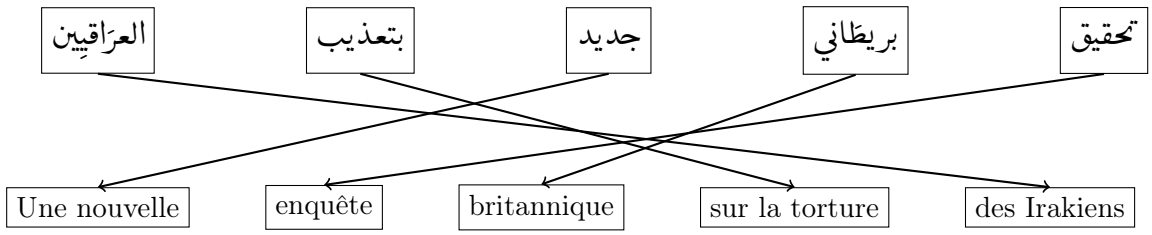


FIGURE 2.9 – Exemple d'alignement entre une phrase arabe et sa traduction française montrant la complexité de réordonnancement des segments.

En outre, l'ordre des mots qui est relativement libre en arabe conduit à des cas où plusieurs phrases peuvent avoir la même traduction. Les quatre phrases arabes du tableau 2.1 ont la même traduction française *le garçon a mangé une pomme*, mais avec un ordre de mots différent. Il est à noter que les quatre possibilités sont valides en arabe.

Ordre des mots	**Phrase arabe**	**Traduction française**
Sujet Verbe Objet (SVO)	الولد اكل التفَاحة	Le garçon a mangé la pomme
Verbe Sujet Objet (VSO)	اكل الولد التفَاحة	A mangé le garçon la pomme
Objet Verbe Sujet (OVS)	التفَاحة اكل الولد	La pomme a mangé le garçon
Verbe Objet Sujet (VOS)	اكل التفَاحة الولد	A mangé la pomme le garçon

TABLE 2.1 – Exemple sur les possibilités d'ordonnancement des mots dans une phrase arabe.

Cette complexité de réordonnancement de segments n'est pas limitée à la langue arabe, il existe d'autres langues qui se caractérisent par le même phénomène, par exemple les langues chinoise et allemande. Plusieurs approches ont été proposées pour résoudre ce problème. [Habash, 2007, Wang *et al.*, 2007] ont proposé une phase de normalisation de données en utilisant des règles de réordonnancement syntaxique avant l'alignement des données parallèles. Une autre approche basée sur les corpus arborés, *treebank*, a été proposée par [Rasooli et Collins, 2019]. Ces corpus sont utilisés pour annoter la structure syntaxique ou sémantique des phrases. Dans leur travail, les auteurs ont utilisé des corpus arborés en étiquetant morpho-syntaxiquement les phrases. Leur travail était basé sur l'étude de la dépendance de l'analyse syntaxique entre les phrases de la langue source et celles de la langue cible. Leur idée consiste à réorganiser les arbres des phrases sources pour les rendre similaires à ceux de la langue cible avant la phase de l'apprentissage. Les auteurs n'ont pas rapporté l'effet de leur méthode de réordonnancement sur la traduction automatique. Il est intéressant d'exploiter cette piste et d'étudier la méthode proposée pour la traduction automatique des langues avec un ordre de mots libre comme l'arabe.

Comme nous l'avons déjà mentionné dans la section 1.3, la langue arabe est riche morphologiquement, un seul mot peut être traduit en plusieurs mots dans une autre langue. Ceci est principalement dû au grand nombre de préfixes et de suffixes qui existent dans la langue arabe et qui sont généralement collés aux mots. Cela entraîne un vocabulaire plus large et, par conséquent, une augmentation potentielle du taux de mots hors vocabulaire. Pour résoudre ce problème, de nombreux travaux précédents indiquent qu'une simple analyse morphologique des mots arabes est utile et ont montré de bons résultats pour la traduction automatique [Ceausu et Tufis, 2012].

Comme dans les systèmes de reconnaissance automatique de la parole, la plupart des travaux proposés dans le domaine de la traduction automatique concerne la traduction de l'arabe standard. Peu de travaux sont disponibles pour le dialecte. Ceci est particulièrement dû au manque

de ressources et au fait que les dialectes sont principalement parlés. Les travaux sur la traduction des dialectes se sont focalisés sur la traduction vers l'arabe standard ou l'anglais. Les premières approches pour la traduction des dialectes arabes étaient basées sur l'application d'un ensemble de règles linguistiques permettant de traduire les mots dialectaux selon le contexte et de les réordonner [Abo-Bakr *et al.*, 2008, Salloum et Habash, 2012, Mohamed *et al.*, 2012, Tachicart et Bouzoubaa, 2014]. Toutefois, le développement de ces systèmes de traduction automatique à base de règles est très coûteux et ne garantit pas de bonnes traductions.

L'utilisation des approches à base de règles était justifiée par le manque de ressources parallèles. Pour remédier à ce problème, [Meftouh *et al.*, 2015] ont proposé un corpus multilingue PADIC contenant plus de 6000 phrases dans plusieurs dialectes avec leur traduction en arabe standard. Une étude détaillée sur la faisabilité de l'utilisation de l'approche statistique de traduction automatique pour la traduction des dialectes a été réalisée [Meftouh *et al.*, 2015, Meftouh *et al.*, 2018]. Ce corpus a été utilisé dans plusieurs travaux de recherche dont celui proposé par [Baniata *et al.*, 2018] où les auteurs ont étudié l'impact de l'apprentissage multitâches sur la traduction des dialectes.

Plusieurs langues ont un impact sur les dialectes arabes, notamment le français et l'anglais, ce qui entraîne l'emprunt de plusieurs mots à partir de ces langues. Cet emprunt influence l'écrit et le parler des dialectes dans le monde arabe. Au niveau lexical, les personnes préfèrent d'utiliser les caractères latins pour écrire leur messages, c'est ce qu'on appelle l'arabizi. Tandis qu'au niveau du parler, les arabes ont tendance à utiliser plus de deux langues dans leurs discours (le dialecte et une autre langue étrangère), ce phénomène est appelé l'alternance codique, ou le *code-switching*. Ce dernier sera discuté en détail dans le chapitre 6.

L'arabizi est un système d'écriture non standard qui est basé sur les caractères latins pour écrire les dialectes arabes. Il est largement utilisé dans le contexte des communications sur les réseaux sociaux et les SMS. Bien que l'Arabizi n'obéisse à aucune règle linguistique, il est largement utilisé et une grande quantité de données disponibles sur les sites communautaires est basée sur ce système d'écriture. Par conséquent, de nouveaux défis sont imposés en particulier pour la traduction automatique. Les approches traitant l'arabizi dans le domaine de la traduction automatique sont basées sur l'idée de convertir les séquences de mots écrites en script latin vers le script arabe avant leur traduction vers une langue [van der Wees *et al.*, 2016]. Le problème de conversion des séquences de mots arabizi vers le script arabe peut être vu comme un problème de traduction où la langue source est l'ensemble des séquences de mots écrites en script latin et la langue cible est l'équivalent de ces séquences en script arabe [May, 2014]. L'arabizi rend difficile la détection de la langue utilisée dans le texte. En effet, dans la même séquence de mots, on peut trouver de la langue française, de la langue anglaise et du dialecte (voir l'exemple du tableau 2.2). Dans nos travaux de thèse sur le dialecte, nous avons décidé de considérer seulement les séquences de mots écrites en scripte arabe.

Dialecte	**Français**	**Dialecte**	**Français**	**Dialecte**	**Anglais**
tbarak allah 3lik	merci pour la vidéo	3djbni	les boucles d'oreilles	mnin khdhitihom	thanks.

TABLE 2.2 – Exemple d'une phrase du dialecte marocain écrite en arabizi avec le phénomène du codeswitching. Trois langues sont utilisée dans cette phrase : le français, l'anglais et le dialecte. La traduction de la phrase *Que Dieu vous bénisse, merci pour cette vidéo, j'ai bien aimé les boucles d'oreilles vous les avez achetées où. Merci.*

2.6 Conclusion et discussion

Nous avons présenté dans ce chapitre les principes de base de la traduction automatique ainsi que les principales techniques et approches proposées dans le domaine. Nous avons abordé en détail l'approche statistique à base de segments et l'approche neuronale. Ces deux approches seront comparées par la suite dans nos travaux de thèse.

L'approche statistique de la traduction automatique s'inspire de l'idée de la modélisation statistique des systèmes de reconnaissance automatique de la parole. Elle est basée sur deux modèles : le modèle de traduction qui permet de trouver pour un mot ou une expression dans la langue source sa traduction dans la langue cible et le modèle de langage qui assure que la traduction est bien écrite dans la langue cible. Ces deux modèles sont exploités conjointement par le décodeur qui explore l'espace de recherche dans le but de trouver la meilleure traduction. Pour les langues ayant une morphologie complexe ou un ordre de mots relativement libre comme la langue arabe, ces deux modèles ne suffisent pas. Des modèles complémentaires sont intégrés dans le processus de décodage pour capturer d'autres caractéristiques linguistiques, à savoir le modèle de réordonnancement et le critère de longueur de la traduction (pénalité des mots).

L'approche neuronale vient remplacer tous les modèles utilisés dans l'approche statistique par un seul modèle à base de réseau de neurones. Ce dernier est généralement de type séquence-à-séquence où la phrase source est projetée dans un espace multidimensionnel en utilisant un encodeur. Le décodeur génère la phrase cible mot à mot à partir de la représentation intermédiaire de la phrase source. En pratique, bien que le modèle neuronal soit capable de capturer les règles linguistiques de la langue cible, un modèle de langage peut être intégré dans le processus de traduction pour améliorer les résultats.

Dans nos travaux de thèse, nous travaillons sur la traduction de l'arabe vers la langue anglaise. Les défis que présentent la langue arabe ainsi que les travaux réalisés dans ce domaine ont été présentés à la fin de ce chapitre.

Deuxième partie

Contributions

Chapitre 3

ALASR : un système de reconnaissance automatique de la parole arabe

Sommaire

Les approches statistiques pour le développement des systèmes de reconnaissance automatique de la parole sont indépendantes de la langue ; néanmoins, il faut prendre en considération ses caractéristiques afin d'apprendre d'une manière efficace les différents modèles.

Dans ce chapitre, nous présentons notre *recette* de développement d'un système de reconnaissance automatique de la parole pour l'arabe [Menacer *et al.*, 2017b]. Ce système sera appelé,

dans ce qui suit, ALASR pour *Arabic Loria Automatic Speech Recognition system*. Le développement de ce système est un premier pas pour notre objectif final qui consiste à traduire la parole arabe.

Nous commençons par une description générale de la langue arabe et de ses caractéristiques ainsi que les défis que notre système doit surmonter pour une meilleure reconnaissance de cette dernière. Par la suite, nous allons présenter nos approches pour l'apprentissage des différents modèles, à savoir : le modèle acoustique et le modèle de langage. Pour finir, on va présenter les tests effectués ainsi que les résultats obtenus tout en comparant notre système aux systèmes état de l'art en termes du WER.

3.1 La langue arabe

Selon Babel, le leader dans le domaine de l'apprentissage des langues en ligne, l'arabe est classée parmi les six principales langues du monde avec plus de 300 millions de locuteurs natifs. C'est la langue officielle des 22 pays qui forment la ligue arabe et l'une des six langues officielles des nations unies. Bien qu'elle soit utilisée par les personnes qui vivent principalement dans la région qui s'étend du Moyen-Orient à l'Afrique du nord, elle est largement parlée dans des pays comme la Somalie et c'est la langue des écrits sacrés des musulmans à travers le monde.

L'arabe appartient au chef de file des langues afro-asiatiques précisément au groupe des langues sémitiques qui comprent également l'hébreu, l'araméen, l'akkadien et l'amharique, la langue principale de l'Éthiopie, etc. Certaines langues sémitiques ont disparu ou leur usage s'est fortement réduit en raison de la propagation de la langue arabe qui est devenue la langue sémitique la plus parlée de nos jours.

La langue arabe est unique car elle comprend techniquement plusieurs variantes (arabe classique, arabe standard et dialectes), mais elle est généralement classée comme une seule langue. Une grande partie de ce qui est connu de l'arabe écrit, classique ou ancien provient des événements enregistrés dans le Coran et les textes religieux.

Ces dernières années, le terme arabe standard moderne (MSA) est apparu. Cette variante de langue est identique à l'arabe classique du Coran, à l'exception de l'ajout des mots modernes et de quelques différences dans les constructions grammaticales. Bien que l'arabe standard soit couramment utilisée dans les lieux de travail, par les gouvernements, les médias et dans les établissements d'enseignement, elle n'est plus, pour autant, utilisée dans les discussions quotidiennes. L'arabe parlé a subi des modifications au cours du temps dues principalement aux évènements historiques dont les colonisations. Il s'est imprégné des langues natives des régions, d'où l'apparition de plusieurs variantes qu'on appelle les dialectes.

L'arabe dialectal ou l'arabe parlé est la langue maternelle de la plupart des personnes d'origine arabe. C'est cette variante de langue qui est utilisée dans les discussions informelles et les conversations quotidiennes. Il existe principalement trois facteurs qui ont mené au développement de l'arabe dialectal : l'arabisation, les colonisations et l'influence culturelle des autres pays.

> L'arabisation est le processus qui a conduit la population des régions locales à adopter la langue et la culture arabe lors de l'expansion de l'islam. Durant cette époque, les autochtones ont découvert l'islam qui était transmis principalement grâce à la langue arabe. Ils ont dû simplifier quelques mots pour faciliter leur prononciation ; citons à titre d'exemple le mot d'origine arabe فأر *fʾar* (souris) qui est altéré par les Nord-Africains et devient فار *fār* à cause de la difficulté de prononciation de *Alif* أ *ʾa* entre deux consonnes.

En contrepartie, la colonisation a enrichi les dialectes par le phénomène de l'emprunt. Ce dernier consiste à intégrer des mots du vocabulaire du colonisateur dans le vocabulaire de l'arabe parlé. Ces nouveaux mots soit ils ont gardé la même prononciation que dans la langue originale, par exemple le mot espagnol *bicicleta* (bicyclette) dans le dialecte égyptien /b i s i k l e t a/, ou soit on a altéré leur prononciation comme pour le mot français *table* dont la prononciation est devenue /t‘ a b l a/ dans le dialecte algérien.

L'influence culturelle est due principalement aux mouvements migratoires, au commerce, et plus récemment aux médias et aux réseaux sociaux. Tous ces paramètres ont permis aux dialectes de s'enrichir encore plus, d'où l'apparition de nouveaux mots tels que جمجمولي *ğmğmwly* (Mettez moi des "j'aime") qui est apparu suite à l'émergence des réseaux sociaux.

L'utilisation du dialecte est devenue très importante au point où certaines langues formelles ont été influencées par ce dernier. On peut citer à titre d'exemple les deux mots شوية *šwyh* (un peu, chouïa) et بزّاف *bzāf* (trop, bezef), utilisés dans le dialecte algérien, qui viennent d'être introduits dans le dictionnaire français dû au mouvement migratoire des populations nord-africaines en France.

Géographiquement, il existe quatre grandes classes de dialectes selon les régions d'utilisation comme le montre la figure 3.1.

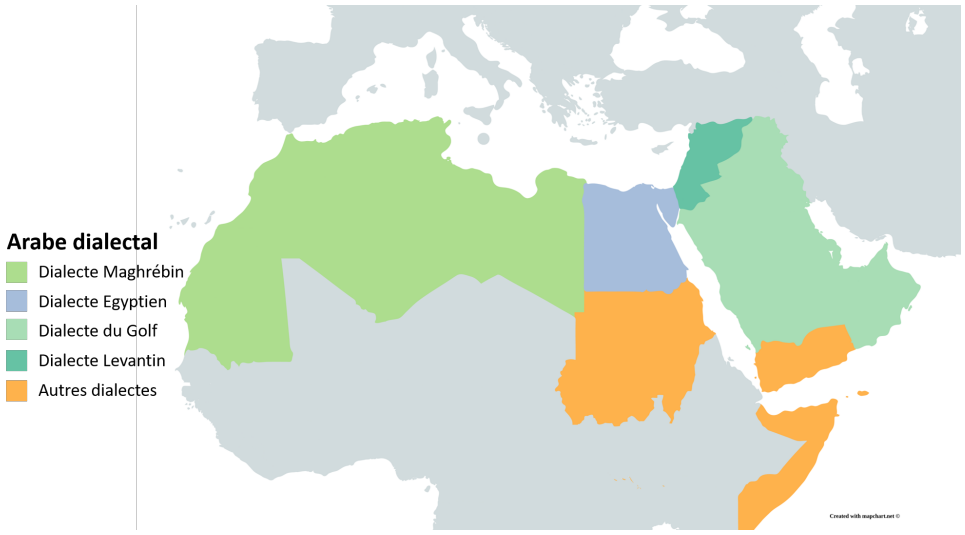


FIGURE 3.1 – Classification des dialectes dans le monde arabe.

L'égyptien est le dialecte le plus parlé avec plus de 100 millions de locuteurs ; il est principalement utilisé en Égypte. Cette variante est l'un des dialectes les plus compréhensibles par les autres arabes grâce aux médias égyptiens qui ont un impact important dans le monde arabe. L'arabe égyptien est fortement influencé par la langue copte, qui était la langue maternelle en Égypte avant la conquête islamique. Aujourd'hui, il contient également des influences linguistiques de plusieurs langues formelles notamment le français, l'italien, le turc et l'anglais.

L'arabe **maghrébin** compte plus de 90 millions de locuteurs répartis entre le Maroc, l'Algérie, la Tunisie, la Libye et le Mauritanie. Cette variante de l'arabe est la plus difficile à comprendre par les autres arabes à cause de nombreuses différences qui existent avec l'arabe standard. Le dialecte maghrébin continue de croître en importance en intégrant de

nouveaux concepts dans la langue. Cela se produit surtout avec l'intégration de nouveaux mots français et/ou anglais et l'adaptation des règles de l'arabe standard pour simplifier leur utilisation et intégration.

Le dialecte du **Golf** est l'arabe parlé dans la péninsule arabique par plus de 36 millions de locuteurs. Cette zone comprend les pays suivants Koweït, Bahreïn, Émirats Arabes Unis, Qatar, l'Arabie Saoudite et Oman. L'arabe du Golfe est une collection de plusieurs dialectes si proches qu'ils peuvent être tous regroupés dans une seule classe. En revanche, il peut y avoir une différence au niveau du vocabulaire, de la grammaire et surtout de l'accent. Cette différence s'accentue avec la distance géographique entre les régions. Un bon exemple est la version de l'arabe du Golf parlé par des personnes au Koweït et au Qatar. Ces deux dialectes peuvent être si différents que les locuteurs peuvent avoir du mal à se comprendre. Cela est principalement dû à l'accent de chaque dialecte.

L'arabe **levantin** est le plus proche de l'arabe standard. En effet, les locuteurs de ce dialecte, qui sont estimés à plus de 32 millions, restent fidèles à l'arabe standard quand il s'agit de l'écriture particulièrement sur les réseaux sociaux ; ils utilisent le dialecte seulement pour la communication orale. Il est parlé dans la bande côtière de la méditerranée orientale qui mesure 100 à 200 kilomètres de large. Ce dialecte est le résultat d'un processus de changement de langue de l'araméen (la langue originelle utilisée dans la région) vers l'arabe qui s'est produit à travers des générations de bilinguisme araméen et arabe. Ce changement a commencé à se produire au 7ème siècle.

Une autre classification plus sophistiquée de l'arabe dialectal existe. En effet, chaque classe citée ci-dessus peut être décomposée en plusieurs sous classes dépendant des régions géographiques plus restreintes où les dialectes sont parlés. Si on considère par exemple le dialecte maghrébin, il est composé de plusieurs autres variantes à savoir le dialecte marocain, l'algérien, le tunisien, le libyen et le hassanya ou l'arabe mauritanien. Chacune de ces variantes comprend aussi d'autres sous catégories dépendant de la distance géographique entre les régions, par exemple les personnes de l'est algérien utilisent un dialecte différent de ceux de l'ouest.

Dans ce chapitre, nous allons nous concentrer sur le développement d'un système de reconnaissance automatique de la parole pour l'arabe standard. Un aperçu sur les principaux défis relatifs à la reconnaissance automatique de l'arabe est présenté dans un premier temps. Par la suite, nous présentons nos solutions pour surmonter ces défis.

3.2 Défis pour les systèmes de reconnaissance automatique de la parole arabe

Comme on a vu dans le chapitre 1, la mise en place d'un système de reconnaissance automatique de la parole est basée sur trois modèles, à savoir le modèle acoustique, le modèle de langage et le modèle de prononciation. Pour mieux expliquer les défis qu'on doit surmonter pour développer un système robuste de reconnaissance automatique de la parole arabe, on va présenter les aspects qui caractérisent l'arabe standard selon deux catégories : les aspects acoustiques et les aspects syntaxiques.

3.2.1 Aspects acoustiques

Contrairement à la langue anglaise ou la langue française, l'arabe standard est une langue pauvre en voyelles et riche en consonnes. Il existe seulement trois voyelles dont chacune possède

une forme longue et courte. Les 28 lettres de l'arabe sont utilisées pour les consonnes et les voyelles longues آ ā /a: /, و w /u: / et ي y /i: / ; les voyelles courtes sont représentées par des diacritiques au-dessous/au-dessus des lettres *fat-ha* ـَ /a/ (une petite barre située au-dessus de la lettre), *dham-ma* ـُ /u/ (un petit و w situé au-dessus de la lettre) et *kasra* ـِ /i/ (une petite barre située au dessous de la lettre). Du point de vue linguistique, la prononciation des consonnes suivies par des voyelles longues est plus allongées. Sa durée est de deux mouvements (un mouvement correspond à la flexion ou l'extension d'un doigt). En revanche, le degré d'allongement des consonnes lorsqu'elles sont suivies d'une voyelle longue n'est pas autant remarquable dans les discours courants. Dans le cas où aucun son n'est attribué aux consonnes, on parle du *soukoun* ◦ (un petit cercle au-dessus de la lettre). Le *soukoun* n'est pas une voyelle mais son contraire, il est utilisé lorsqu'on prononce une lettre sans aucune voyelle.

À l'inverse des signes diacritiques des voyelles citées précédemment qui sont utilisées pour définir le sens de chaque mot, il existe deux autres signes qui permettent de donner plus de structure aux discours et dont l'importance consiste à capter l'intention de l'interlocuteur, il s'agit de *chad-da* et *tanween*. *Chad-da* ou gémination est un petit signe qui a la forme de la lettre س s et est située au-dessus de la consonne. Elle est prononcée comme une consonne double en mettant l'accent sur celle-ci. *Tanween* consiste à ajouter le son /n/ à la fin de la voyelle. Il existe trois formes de *tanween* selon les trois voyelles courtes اً /a n/, اٌ /u n/ et اٍ /i n/. Ce genre de diacritiques n'affectent seulement la dernière consonne du mot.

Ces signes diacritiques ont été inventés par *Abou al-Aswad al-Douali*, l'un des premiers grammairiens arabes, au $7^{éme}$ siècle. Sa proposition était faite suite aux erreurs de voyellation lors de la récitation du Coran par les non arabophones. Bien que la voyellation ou *tachkil* en arabe n'est pas obligatoire, les voyelles ont un rôle très important du fait que le mot peut changer de sens si la voyelle n'est pas à sa bonne place. Regardons l'exemple classique de la voyellation du mot كتب *ktb*. Selon la position des voyelles, il peut signifier كَتَبَ *kataba* (il a écrit), كُتُب *kutubn* (des livres) ou كُتِبَ *kutiba* (il a été écrit), etc. et il a 21 possibilités de voyellation avec des sens différents [Vergyri et Kirchhoff, 2004]. Un arabophone peut facilement déduire la voyellation d'un mot selon le contexte et en suivant les règles de grammaire arabe.

L'utilisation des voyelles de nos jours est restreinte aux textes religieux, poétiques et aux ouvrages destinés aux enfants pour faciliter la prononciation des mots. Le problème qui se pose pour les systèmes de reconnaissance automatique de la parole est qu'il est difficile d'apprendre un modèle acoustique efficace pour les voyelles sans savoir leur position dans le signal. Ajoutons à cela le grand nombre de sens qu'on peut avoir pour un seul mot non diacritisé. Dans la section 3.5, nous discutons en détail l'impact de modéliser ou non les voyelles, d'une manière explicite, sur les performances du système de reconnaissance de la parole.

3.2.2 Aspects syntaxiques

En arabe, les mots sont écrits dans un style cursif, ce qui signifie que les lettres sont attachées les unes aux autres. Pour cela, chaque caractère est écrit dans quatre formes différentes selon son emplacement dans le mot (au début, au milieu, à la fin et isolé). Un exemple d'écriture du caractère ح *ḥ* est donné dans le tableau 3.1

Position dans le mot	Début	Milieu	Fin	Isolé
Style d'écriture	حـ	ـحـ	ـح	ح

TABLE 3.1 – Les différentes façons d'écrire le caractère ح

L'orthographe de l'arabe est déroutante du fait que l'écriture de certains mots pourrait être simplifiée en remplaçant une lettre par une autre ou en omettant le symbole de *hamza* ء. Certains exemples de confusion sont illustrés dans le tableau 3.2.

Mot correct	**Remplacé par**	**Explication**	**Traduction**
مدرسة	مدرسه	La lettre ة est remplacée par ه	École
زرَاعي	زراعى	La lettre ي est remplacée par ى	Agricole
إستعمَال	استعمال	Le symbole ء au-dessous de *Alif* ا est supprimé	Utilisation

TABLE 3.2 – Exemples de quelques erreurs d'orthographe.

Le dernier exemple dans le tableau 3.2 est le plus fréquent. Les personnes ont tendance à simplifier l'écriture en supprimant le symbole *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا. Malheureusement, ces erreurs d'écriture existent dans les journaux et/ou les documents officiels. Ainsi, dans un même document, il peut y avoir les deux écritures (correcte et erronée) ce qui conduit à répartir la probabilité d'un mot contenant cette lettre en deux formes d'un même mot. Évidemment, il existe d'autres cas à prendre en considération pour harmoniser l'orthographe dans les documents arabes. Certains d'entre eux sont spécifiques à l'arabe et d'autres sont utilisés dans la majorité des langues naturelles. Ces cas seront discutés en détail dans la section 3.6.

Un autre point qui caractérise la langue arabe est la richesse de sa morphologie. En concaténant des antéfixes, des préfixes, des suffixes et des postfixes aux thèmes *stem*, d'autres mots sont obtenus. Le thème morphologique peut également être décomposé en une racine (généralement une séquence de trois consonnes), un motif de voyelles et, éventuellement, de consonnes supplémentaires. La figure 3.2 montre une forme plus agglutinée à partir de la racine كتب *ktb*. On peut remarquer qu'un seul mot en arabe ليكَاتبوهم *lykātbwhm* peut correspondre à toute une phrase en français *pour qu'ils leurs écrivent*.

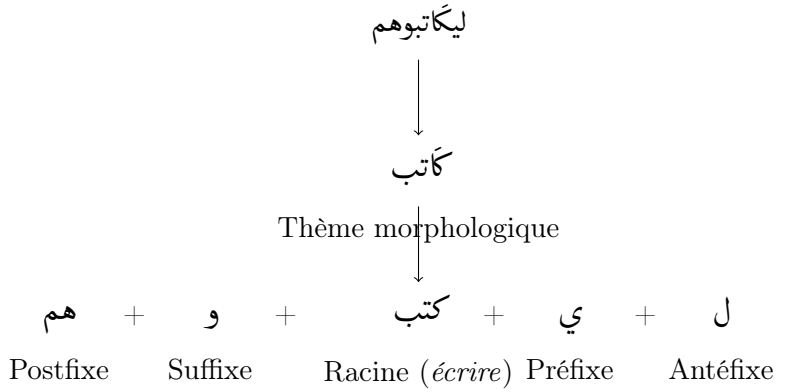


FIGURE 3.2 – Exemple d'une forme agglutinée à partir de la racine كتب (écrire).

La richesse de la morphologie de l'arabe impacte directement l'apprentissage du modèle de langage. En effet, pour un même nombre d'entrées dans le vocabulaire utilisé pour apprendre le modèle de langage, le taux de mots hors vocabulaire en arabe est toujours plus élevé par rapport à l'anglais. Cela impacte également le WER vu que chaque mot hors vocabulaire ne sera pas reconnu par le système de reconnaissance de la parole et qu'il entraîne environ 1,2 erreurs dans les calcul du score final.

Un autre point qui impacte le vocabulaire de l'arabe est la présence des voyelles dans le texte. Bien que les diacritiques jouent un rôle important pour le modèle acoustique, leur utilisation pour apprendre le modèle de langage peut avoir un impact négatif sur les performances du système. Dans ce cas, au lieu d'utiliser une seule forme non voyellée pour un mot donné, on doit en utiliser plusieurs voyellées. En reprenant l'exemple du mot non voyellé كتب *ktb*, il peut y avoir 21 entrées dans le vocabulaire pour représenter ce mot (كَتَبَ *kataba*, كُتُبٌ *kutubun*, كُتِبَ *kutiba*,etc.). Cela implique l'utilisation d'une taille plus importante du vocabulaire afin de réduire le taux de mots hors vocabulaire et améliorer, en conséquence, le WER. Dans les sections suivantes, nous présentons nos approches pour apprendre les différents modèles en prenant en considération les caractéristiques de la langue arabe.

3.3 Données utilisées

L'apprentissage du modèle acoustique et du modèle de langage est basé respectivement sur des corpus oraux et des corpus textuels. Un corpus oral est une collection de fichiers audio avec leurs transcriptions textuelles, quant au corpus textuel, c'est une collection de séquence de mots de la langue à reconnaître. Dans ce qui suit, nous présentons les données utilisées pour l'apprentissage et l'évaluation des différents modèles.

3.3.1 Données orales

L'apprentissage et l'évaluation de notre modèle acoustique sont effectués à partir de deux corpus oraux NEMLAR [Yaseen *et al.*, 2006] et NetDC [Choukri *et al.*, 2004]. Ces deux corpus sont composés de 63 heures de paroles provenant de journaux télévisés.

NEMLAR est le plus grand corpus avec 40 heures de parole enregistrées par 259 locuteurs (80% hommes et 20% femmes) à partir de quatre stations radio différentes. La durée moyenne de chaque émission radio varie entre 25 et 30 minutes. La transcription des fichiers audio comprend différents niveaux d'annotation à savoir la transcription orthographique de la parole, les entités nommées, les locuteurs, les marqueurs de segments, le bruit de fond et la musique.

NetDC est composé de 37 émissions diffusées par Radio Orient d'une durée moyenne de 35 minutes par émission. Le nombre de locuteurs identifiés dans le corpus est 90. Comme dans le corpus NEMLAR, la transcription orthographique contient des informations sur les locuteurs, les sujets abordés, les marqueurs de segments et le bruit.

Nous avons divisé les deux corpus en trois parties comme présenté dans le tableau 3.3.

Corpus	Apprentissage	Validation	Test	Total
NEMLAR	33 (83%)	3 (08%)	3 (9%)	40
NetDC	19 (82%)	3 (10%)	2 (8%)	23
Total	52 (83%)	6 (09%)	5 (8%)	63

TABLE 3.3 – Statistiques sur les données acoustiques (en heures).

3.3.2 Données textuelles

La langue arabe est une langue riche en ressources textuelles monolingues. Elle est classée parmi les dix premières langues sur Internet [9], ce qui facilite l'accès aux données textuelles à partir du web. Deux corpus ont été utilisés dans notre système pour apprendre et évaluer le modèle de langage : Gigaword [Parker *et al.*, 2011] et la transcription des données acoustiques. Gigaword est un large corpus avec plus d'un milliard de mots. Il a été collecté à partir de neuf sources d'informations entre 2003 et 2010. Le tableau 3.4 donne quelques statistiques sur ces deux corpus.

Corpus	Nombre de mots	Nombre de mots uniques
GigaWord	1 000 000 k	3 897 k
Transcription	315 k	38 k

TABLE 3.4 – Statistiques sur les données textuelles.

3.4 Modélisation acoustique

Le signal de la parole comprend plusieurs informations qui ne sont pas forcément nécessaires pour apprendre le modèle acoustique. Pour cela, la première étape consiste à en extraire des vecteurs numériques comportant seulement les informations représentatives du message linguistique. Ces vecteurs seront utilisés par la suite pour trouver les unités acoustiques correspondant à notre signal. Dans notre cas, nous avons opté pour des paramètres MFCC [Davis et Mermelstein, 1980] avec leurs dérivées première et seconde, soit des vecteurs acoustiques de dimension 39.

Notre modèle acoustique est un modèle à base de phonèmes ; pour chaque vecteur numérique on essaie de prédire le phonème qui lui correspond au mieux. Pour l'arabe, nous avons considéré 34 phonèmes (28 consonnes et 6 voyelles), et une unité acoustique non linguistique (*le silence*). Pour représenter ces phonèmes, nous avons utilisé la notation phonétique SAMPA. SAMPA a été originalement conçu pour seulement 6 langues européennes en 1989 et il a été étendu pour couvrir plus de 27 langues en fin 2015. La notation phonétique SAMPA pour l'arabe est constituée de 38 symboles [10], quatre d'entre eux n'ont pas été pris en considération dans notre modélisation acoustique car ils représentent des sons qui ne sont pas utilisés en arabe standard (voir tableau 3.5).

9. source : https://www.internetworldstats.com/stats7.htm
10. Vous pour y accéder à partir de cette page web.

Phonème	API	Exemple d'utilisation	Traduction
/g/ dialectes égyptien	/g/	قميل *vmyl* /g a m i: l/	Beau
/v/ dialecte maghrébin	/v/	مناّرفي *mnārfy* /m n a: r v i: /	Agité/Énervé
/p/ dialecte maghrébin	/p/	پرَاپلي *prāply* /p a r a p l i/	Parapluie
/l‘/ gémination de ل	/l/	اللّه *āl-lah* /? a l‘ h/	Dieu

TABLE 3.5 – Quelques phonèmes SAMPA non pris en considération dans notre système ALASR.

Notre modèle acoustique est basé sur les réseaux de neurones qui sont entraînés pour estimer la probabilité d'associer chaque trame du signal [11] à une unité acoustique. Pour apprendre ce genre de modèles, on doit disposer de données étiquetées de la forme (trame, unité acoustique). En pratique et en raison de l'absence de ce genre de données, un processus automatique appelé alignement forcé est utilisé. Il consiste à entraîner un modèle acoustique de base (généralement un modèle GMM-HMM) et de l'utiliser pour aligner le signal avec les unités acoustiques. L'alignement entre les trames acoustiques et les unités de la modélisation peut être définit au niveau de phonèmes, de caractère ou de mots, etc.

Le réseau de neurones est grandement affecté par le modèle de base utilisé pour l'alignement forcé. Peu importe le nombre d'itération qu'on exécute, le type de fonction de coût qu'on utilise ou le taux d'apprentissage qu'on utilise pour mettre à jour les paramètres du réseau de neurones, entraîner ce dernier avec un mauvais alignement conduit toujours à un mauvais modèle acoustique. Pour cette raison, la première étape est d'entraîner un bon modèle acoustique de base pour aligner nos données et de les utiliser, ensuite, pour apprendre notre réseau de neurones.

3.4.1 Modèles GMM-HMM

Le modèle acoustique utilisé pour aligner nos données d'apprentissage est un modèle de Markov caché avec des densités multigaussiennes GMM-HMM (voir section 1.1.2). En partant de l'idée que les données d'apprentissage ne permettent pas de représenter toute la variabilité entre différents locuteurs et différentes conditions d'enregistrement, nous avons entraîné une série de 4 modèles *triphone* en appliquant des transformations linéaires afin de mieux affiner le modèle acoustique (voir figure 3.3).

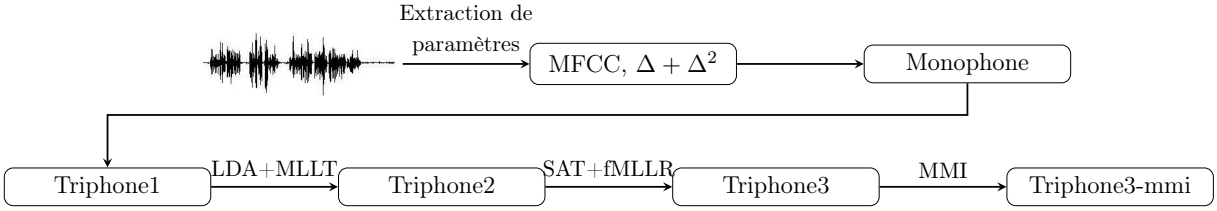


FIGURE 3.3 – L'hiérarchie des modèles GMM-HMM entraînés.

Modèle *monophone.* L'apprentissage des modèles *triphones* commence par entraîner un modèle *monophone* où l'unité acoustique est le phonème. L'objectif à ce niveau est d'entraîner un modèle de base avec peu de paramètres pour initialiser l'apprentissage des modèles *triphones*. Dans ces derniers, le contexte de chaque phonème sera pris en consi-

11. Une trame du signal est représentée par un vecteur acoustique MFCC.

dération d'où l'unité acoustique de base sera un phonème dans son contexte gauche et droit.

Modèle *triphone1*. L'apprentissage de ce modèle est basé sur les observations acoustiques MFCC, $\Delta+\Delta^2$ (dérivées première et seconde).

Modèle *triphone2*. Le deuxième modèle *triphone2* est entraîné en réduisant la dimension des paramètres acoustiques en appliquant une analyse discriminante linéaire, *Linear Discriminant Analysis (LDA)* [Haeb-Umbach et Ney, 1992] ; cette dernière est une projection sur les composantes les plus discriminantes. L'analyse LDA est suivie par une transformation linéaire avec maximisation de la vraisemblance, *Maximum Likelihood Linear Transform (MLLT)* [Gopinath, 1998]. Il s'agit d'une transformation orthogonale des paramètres acoustiques afin de les bien modéliser avec des gaussiennes à matrice de covariance diagonale [Rath *et al.*, 2013].

Modèle *triphone3*. Les deux transformations LDA+MLLT sont indépendantes du locuteur, il existe d'autres transformations qui sont utilisées pour l'apprentissage adaptatif, *speaker Adaptive Training (SAT)* [Anastasakos *et al.*, 1996] et *feature-space Maximum Likelihood Linear Regerssion (fMLLR)* [Gales, 1998]. Cette adaptation vise à rapprocher les paramètres acoustiques initiaux (ceux de l'ensemble d'apprentissage) et cibles (ceux de l'ensemble de validation) par une transformation linéaire. Dans l'apprentissage adaptatif SAT, le modèle acoustique est entraîné selon le processus itératif illustré dans la figure 3.4. À chaque itération de ce processus, un nouveau modèle est construit en se basant sur les données d'apprentissage adaptées.

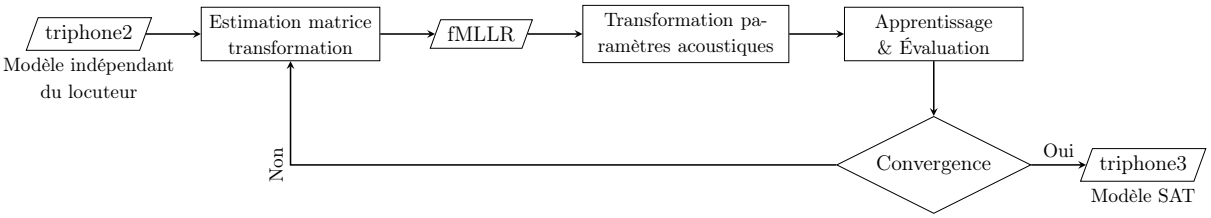


FIGURE 3.4 – Processus d'apprentissage adaptatif SAT.

Modèle *triphone3-mmi*. Pour apprendre les modèles triphones précédents, l'estimateur du maximum de vraisemblance est utilisé comme fonction objective. L'un des principaux problèmes de cet estimateur est qu'il permet au modèle de s'adapter aux données d'apprentissage, et élimine sa capacité à discriminer. Pour y remédier, une approche alternative avec un critère d'apprentissage discriminant est utilisé : *Maximum Mutual Information (MMI)* [Bahl *et al.*, 1986] pour apprendre ce dernier modèle. Dans cette approche, les paramètres du modèle acoustique sont estimés pour maximiser l'information mutuelle entre une observation acoustique $O = o_1 o_2 \dots o_T$ et la séquence de mots $W_r = w_1 w_2 \dots w_l$ qui lui correspond. En pratique, ce critère est calculé en se basant sur les n-meilleures séquences de mots générées par le système de reconnaissance de la parole, ce qui permet non seulement de rendre l'hypothèse correcte plus probable, mais aussi de diminuer la probabilité des autres hypothèses qui ne sont pas correctes.

Le derniers modèle *triphone3* dispose de 4 264 densités multigaussiennes (rappelons que ce nombre correspond aux feuilles de l'arbre de décision présentée dans la section 1.1.2) ; ainsi le nombre total de gaussiennes est de 100 119 ce qui fait en moyenne 23,5 gaussiennes par densité

pour estimer la probabilité de générer une observation acoustique par chaque état du modèle de Markov caché.

Il est à noter que les paramètres acoustiques subissent d'un alignement forcé après l'apprentissage de chaque modèle triphone. L'alignement obtenu avec le meilleur modèle GMM-HMM est celui utilisé pour apprendre notre réseau de neurones.

3.4.2 Modèles à base de réseaux de neurones, *DNN-HMM*

Contrairement aux modèles GMM-HMM où la probabilité d'émission des observations acoustiques pour chaque état est estimée par une somme de lois gaussiennes, dans les modèles DNN-HMM cette probabilité est estimée par le réseau de neurones. Le réseau de neurones prend en entrée un vecteur de paramètres acoustiques et prédit en sortie l'unité acoustique (le triphone dans notre cas) qui correspond au mieux à l'entrée.

Nous avons entraîné deux modèles neuronaux avec deux architectures différentes : le premier modèle est un perceptron multicouche, *multilayer perceptron (MLP)*, et le deuxième est un réseau de neurones à retard temporel, *Time Delay Neural Network (TDNN)*.

Modèle à base de perceptron multicouche

Le perceptron multicouche, ou réseau neuronal à propagation avant, *feedforward*, est un réseau de neurones organisé en plusieurs couches où l'information est propagée de la première couche, celle de l'entrée, jusqu'à la dernière, celle de la sortie.

Notre architecture est basée sur 6 couches cachées de 2 048 neurones chacune. La couche en entrée est composée de 440 neurones représentant la concaténation de 11 observations acoustiques. En effet, pour chaque instant t, le réseau de neurones prend en entrée la concaténation des $[t-5; t+5]$ vecteurs fMLLR de dimension 40. La couche de sortie est composée de 4 264 neurones représentant le nombre total de densités multigausiennes de la modélisation GMM-HMM. Le nombre total de paramètres à estimer est de 30,6 millions.

Pour apprendre les paramètres du réseau de neurones, nous avons comparé trois critères d'apprentissage : l'entropie croisée, *Minimum Phone Error* (MPE), *state Minimum Bayes Risk* (sMBR). Rappelons que les deux critères MPE et sMBR fonctionnent au niveau des séquences de mots contrairement à l'entropie croisée qui fonctionne aux niveaux des trames.

En ce qui concerne la technique d'estimation des paramètres du perceptron multicouche, nous avons utilisé l'algorithme de la descente de gradient stochastique où l'estimation est faite sur des *mini-batch* de 256 segments avec un *learning rate* de 0,008.

Modèle à base du réseau de neurones à retard temporel *(TDNN)*

Les TDNN sont des réseaux à base de plusieurs couches qui se distinguent par rapport aux modèles de perceptron multicouche par leur architecture hiérarchique leur permettant de mieux capturer le contexte. Cette architecture a été proposée initialement par [Waibel *et al.*, 1989] pour la reconnaissance de phonèmes et elle a été adaptée par la suite pour la reconnaissance automatique de la parole continue [Peddinti *et al.*, 2015]. Le TDNN dispose de fenêtres de contexte qui sont utilisées pour balayer, d'une part, le signal acoustique au niveau de la couche d'entrée et, d'autre part, les sorties intermédiaires des couches cachées. La couche en entrée traite les paramètres acoustiques avec un contexte étroit tandis que des contextes plus larges seront traités par les couches supérieures. Cela est différent de l'utilisation d'une large fenêtre contextuelle de vecteurs de paramètres acoustiques dans la couche d'entrée, comme nous avons fait avec le modèle à base de perceptron multicouche où pour chaque instant t, on a pris la concaténation

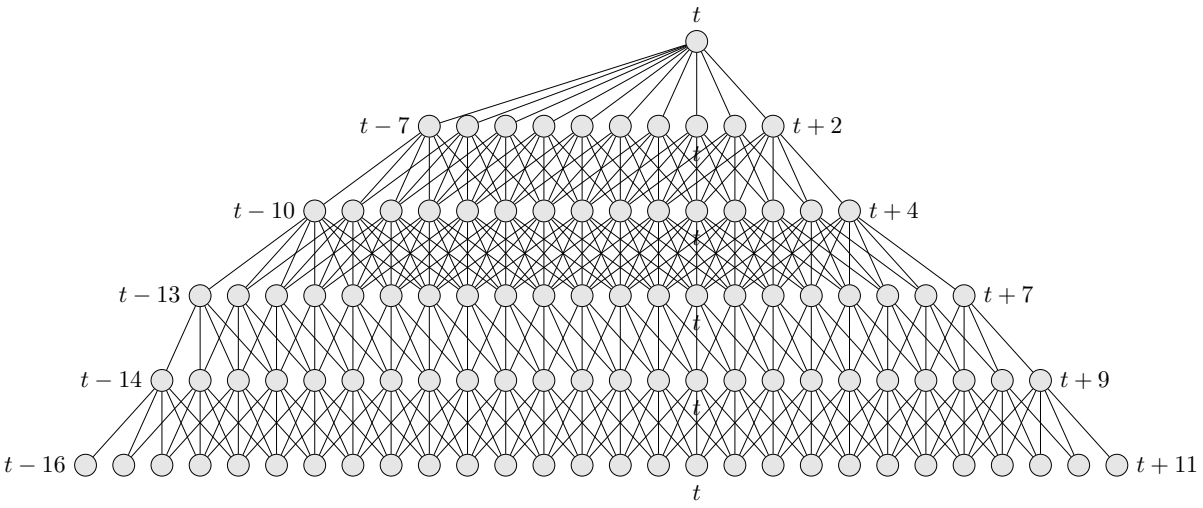


FIGURE 3.5 – Modèle acoustique à base de TDNN.

de 11 vecteurs acoustiques $[t-5; t+5]$. Dans le modèle à base de TDNN, chaque couche cachée extrait des informations abstraites à partir du signal à des niveaux différents, ce qui permet aux couches supérieures d'apprendre des relations temporelles plus larges.

La figure 3.5 montre l'architecture utilisée dans notre système pour apprendre le modèle acoustique. La première couche prend en entrée la concaténation de 5 vecteurs acoustiques $[t-2; t+2]$. La deuxième couche s'appuie sur une fenêtre de contexte de 4 vecteurs représentant la sortie de la couche une ($[t-1; t+2]$). Les fenêtres de contexte utilisées dans chaque couche cachée sont illustrées dans le tableau 3.6.

Layer	1	2	3	4	5	6
Contexte	[-2,+2]	[-1,+2]	[-3,+3]	[-3,+2]	[-7,+2]	{0}
Contexte avec sous-échantillonnage	[-2,+2]	{-1,+2}	{-3,+3}	{-3,+2}	{-7,+2}	{0}

TABLE 3.6 – Fenêtres de contexte pour chaque couche du modèle TDNN.

Avec cette architecture, on peut remarquer que la dernière couche a une relation indirecte avec la couche d'entrée qui traite les paramètres acoustiques via la fenêtre de contexte $[t-16; t+11]$. Cette méthode pour constituer les fenêtres de contexte augmente le nombre de paramètres à estimer à cause du grand chevauchement entre les fenêtres de contexte. Ce problème est résolu en restreignant le contexte à seulement deux vecteurs, on parle du sous-échantillonnage du contexte [Peddinti *et al.*, 2015]. En effet, la notation $\{-1, +2\}$ du tableau 3.6 signifie que la deuxième couche se base sur 2 vecteurs ($t-1$ et $t+2$) au lieu de quatre vecteurs pour calculer sa sortie. Cette approche est justifiée par le fait que les sorties des neurones voisins sont corrélées donc on peut réduire le contexte aux deux extrémités de la fenêtre de contexte [Peddinti *et al.*, 2015].

Le modèle basé sur le réseau TDNN est entraîné, dans notre cas, avec une fonction objective au niveau de séquences, à savoir le critère sMBR. Les paramètres sont estimés avec l'algorithme de la descente de gradient stochastique sur des *mini-batch* de 128 segments et un *learning rate* de 0,0002.

3.5 Modélisation de la prononciation

Dans un système de reconnaissance automatique de la parole, on a besoin d'une ressource lexicale qui fait le lien entre les unités acoustiques et les mots, il s'agit du lexique de prononciation. Deux lexiques sont nécessaires : un lexique utilisé lors de la phase d'apprentissage du modèle acoustique, il contient les prononciations de tous les mots qui se trouvent dans la transcription textuelle des données acoustiques du corpus d'apprentissage. Un lexique de reconnaissance qui contient tous les mots qui pourront être reconnus par notre système.

La construction de ces lexiques est basée sur un processus automatique où toutes les prononciations possibles de chaque mot doivent être générées. Le problème majeur de la langue arabe est que les voyelles ne sont pas explicitement écrites dans le texte, ce qui rend difficile la génération de prononciation de mots. Deux solutions ont été mises en oeuvre dans nos évaluations :

— Modéliser implicitement les voyelles : le processus de génération de prononciation est une fonction bijective où chaque mot possède exactement une seule prononciation. Les voyelles ne sont pas restituées dans ce cas, et la prononciation n'est que la décomposition en lettres du mot. Si on reprend l'exemple du mot كتب *ktb*, toutes les prononciations seront représentées par /k t b/. Bien que cette solution soit facile à mettre en place, elle n'assure pas que les voyelles seront apprises implicitement par le modèle acoustique car il est difficile de savoir leur position dans le signal. Le nombre des unités acoustiques dans cette solution est de 28 phonèmes représentant les consonnes de la langue arabe.

— Modéliser explicitement les voyelles : dans cette solution, chaque mot possède au moins une prononciation. Pour avoir toutes les prononciations possibles de n'importe quel mot, les voyelles doivent être restituées dans le texte. Ensuite, la prononciation des mots voyellés est générée avec le même processus dans la modélisation implicite des voyelles. Il existe plusieurs approches qui permettent la restitution des voyelles dans le texte Arabe. [Diehl *et al.*, 2012] ont proposé un outil qui s'appelle MADA et qui permet de générer toutes les prononciations possibles ordonnées par un score de confiance. Leur approche est basée sur l'analyse morphologique de toutes les possibilités de voyellisation pour savoir la forme la plus probable dans un contexte syntaxique. MADA a été utilisé par [Ali *et al.*, 2014] pour construire un lexique de prononciation à partir de mots non voyellés ; pour 526K mots, ils ont obtenu 1,8M variantes de prononciation, ce qui fait en moyenne 3,43 prononciations par mot. Nous avons utilisé ce lexique pour récupérer les différentes variantes de prononciation de tous les mots de notre lexique. Afin d'étudier l'impacte de la modélisation explicite des voyelles courtes, longues et de la gémination sur les performances d'ALASR, nous avons décidé de traiter chaque cas séparément :
  — Modélisation explicite des voyelles courtes : dans ce cas, il n'existe pas de différence entre les voyelles courtes et longues. Le nombre d'unités acoustiques utilisées est de 32 phonèmes (28 consones et 3 voyelles).
  — Modélisation explicite des voyelles longues : pour ces dernières, nous avons décidé d'ajouter trois autres voyelles, ce qui augmente le nombre d'unités acoustiques à 35 phonèmes (28 consones, 3 voyelles courtes et 3 voyelle longues).
  — Modélisation explicite de la gémination : nous avons décidé de modéliser la gémination en doublant le caractère concerné dans la prononciation. Le nombre d'unités acoustiques utilisé est le même que celui de la modélisation explicite des voyelles longues (35 phonèmes).

Notre lexique de reconnaissance est composé des mots les plus fréquents dans les corpus textuels utilisés pour l'apprentissage du modèle de langage. Il représente l'union des mots qui apparaissent plus de 3 fois dans la transcription textuelle des données acoustiques et les 109k mots

les plus fréquents dans le corpus Gigaword. Après avoir utilisé des lexiques externes pour générer les différentes variantes de prononciations de chaque mot, notre lexique de reconnaissance final est composé de 95k mots et de 485k prononciations, ce qui fait en moyenne 5,07 prononciations par mot et cela dans le cas où les voyelles (courtes et longues) sont explicitement modélisées. Dans le cas contraire, chaque mot possède exactement une seule prononciation ce qui fait un lexique de 95k entrées. Pour montrer la différence entre les approches de génération de prononciations, nous donnons dans le tableau 3.7 un exemple d'une entrée de notre lexique qui correspond au mot كرَاس *krās* (cahier) ainsi que les variantes de prononciation selon l'approche utilisée.

Mot	**Prononciation**	**Approche**
كُرَاسٌ *kuraāsun*	/k u r a s u n/	Modélisation explicite des voyelles courtes
كُرَّاسٌ *kurraāsun*	/k u r r a s u n/	Modélisation explicite de la gémination
كُرَّاسٌ *kurraāsun*	/k u r r a : s u n/	Modélisation explicite des voyelles longues
كراس *krās*	/k r s/	Modélisation implicite des voyelles

TABLE 3.7 – Variantes de prononciation du mot كرَاس (cahier) selon les aspects de modélisation de la prononciation.

3.6 Modélisation du langage

L'apprentissage du modèle de langage est basé sur des corpus textuels de la langue à reconnaître, à savoir la langue arabe standard dans notre cas. Ces corpus sont généralement extraits à partir de journaux, de documents officiels, de livres ou même à partir de réseaux sociaux ces dernières années. Quelle que soit la source de nos corpus, il existe toujours des anomalies dans le texte écrit et particulièrement dans la langue arabe. Ces anomalies ont un impact directe sur la modélisation du langage du fait que plusieurs formes d'écriture, qui désignent le même mot, partagent la même probabilité. Pour cette raison, une étape de normalisation des données textuelles est nécessaire pour réduire les conflits d'écriture et donc pour améliorer l'estimation des probabilités des séquences de mots par le modèle de langage.

3.6.1 Normalisation des données

En analysant nos corpus d'apprentissage, nous avons pu établir une liste des erreurs les plus fréquentes. Certaines d'entre elles sont spécifiques à l'arabe et d'autres sont communes à la majorité des langues naturelles. Cette liste ainsi que les solutions proposées sont décrites dans le tableau 3.8.

Vu que le préfixe و *w* (et) est le plus utilisé en arabe, nous avons décidé de le séparer du mot pour réduire la taille du vocabulaire et couvrir plus de mots. En séparant, par exemple, le préfixe du mot ورجَال *wrğāl* (et des hommes), les trois mots و *w* (et), رجَال *rğāl* (hommes) et و رجَال *w rğāl* (et des hommes)) seront reconnus par le système de reconnaissance. En revanche, dans le cas où le préfixe est collé, le mot رجَال *rğāl* (hommes) sera considéré comme un mot hors vocabulaire s'il n'apparaît pas dans un autre contexte.

En plus des opérations listées dans le tableau 3.8, une erreur souvent commise en arabe est la simplification de l'écriture de la lettre أ. Le symbole de *hamza* ء au-dessus/au-dessous de

Contexte	Exemples	Opération
Adresses mails - Chemins URL - Phrases avec du texte non arabe - Diacritiques	و عنوَان هذ الموقع www.arabsummit.org.sa (et le titre de ce site est www. arabsummit.org.sa)	Suppression de la phrase entière
Ponctuations - Diacritiques	انصَارِ => انصار (supporteurs)	Suppression
Les mots étirés	الرجاااال *ālrğāāāāāl* → الرجال *ālrğāl*	Suppression des caractères en double
Le préfixe و *w* (et)	ورجال => و رجال (et des hommes)	Séparation du préfixe et le mot suivant en utilisant l'outil Farassa [Abdelali *et al.*, 2016]
Les autres préfixes ب (avec, dans, par), ف (ensuite), ال (le,la,l'), ك (comme), ل (à, car) et س (exprime le futur)	بِالمتمردين => ب المتمردين (par les rebelles)	Concaténation avec le mot suivant
ة *h*	ازمةقلبية => ازمة قلبية	Insertion d'un espace après le caractère ة *h* (ce caractère est toujours écrit en fin du mot)
Temps	15 :30 → الثالثة وثلاثون دقيقة	Écriture littérale
Nombres	150 => مأة وخمسون	Écriture littérale
Les abréviations	ق . م => قبل الميلَاد (avant JC)	Remplacement par la séquence de mots correspondante (voir annexe A)

TABLE 3.8 – Liste des opérations de normalisation des données textuelles.

la lettre *Alif* ا est souvent supprimé. La même remarque est observable pour le tilde ̃ qui est utilisé dans le cas où le caractère ا est suivi par une voyelle longue (آ). L'approche que nous proposons pour normaliser l'écriture de *hamza* ء est inspirée des techniques qui ont été proposées dans la littérature pour détecter et corriger automatiquement les fautes d'orthographe. C'est un problème commun à toutes les langues naturelles. En arabe, les fautes d'orthographe les plus fréquentes sont celles qui concernent les erreurs d'édition et celles qui concernent les erreurs sémantiques. Le premier type d'erreur se produit lorsqu'un mot est mal orthographié, tandis que pour les erreurs sémantiques, un mot est remplacé par un autre mot correctement orthographié [Alkanhal *et al.*, 2012]. La suppression du symbole *hamza* ء pourrait conduire à un mot mal orthographié (par exemple le mot mal écrit لان *lān* (car), il doit être écrit comme suit لأن *ʾan*) où il pourrait changer le sens du mot (par exemple le mot امَام *āmām* qui pourrait être interprété comme أمَام *ʾmām* (devant) ou إمَام *ʾimām* (imam) selon la position de *hamza* ء). Pour restituer le symbole *hamza* ء et le tilde ̃, nous avons utilisé une procédure basée sur les trois étapes suivantes :

Détection des erreurs deux principales techniques sont utilisées pour détecter les fautes d'orthographe en arabe : la méthode à base de règles [AlShenaifi *et al.*, 2015, Shaalan *et al.*, 2010, Hassan *et al.*, 2014] et la méthode à base de dictionnaire [Attia *et al.*, 2014, Zerrouki *et al.*, 2014, Alkanhal *et al.*, 2012]. Pour la méthode à base de règles, détecter si un mot est mal orthographié ou non dépend de l'analyse morphologique du mot, tandis que dans la technique

basée sur le dictionnaire, la détection dépend d'une grande liste de mots qui couvre les mots les plus fréquemment utilisés dans la langue. Dans notre cas, nous avons opté pour l'utilisation d'une grande liste de mots [Attia *et al.*, 2012] contenant 9,2 millions de mots. Un mot avec la lettre *Alif* ا est correctement orthographié s'il existe dans la liste de mots.

Production d'hypothèses la distance d'édition est la technique la plus utilisée pour produire la liste des mots candidats. Elle mesure la différence entre deux séquences de caractères en calculant le nombre de modifications requises pour transformer un mot en un autre. La correction de l'orthographe du symbole *hamza* ء ne nécessite pas l'utilisation d'une distance d'édition. En effet, la correction est uniquement basée sur la sélection de tous les mots à partir de la liste de mots arabes dont l'orthographe est la même que celle du mot mal écrit à l'exception de *hamza* ء au-dessus/au-dessous de *Alif* ا (أ et إ). Pour clarifier cette démarche, considérons le mot mal orthographié اعمل *ʿāmal*, la liste des candidats contient les deux mots أَعمَل *ʾaʿmal* (je travaille) et إِعمل *ʾiʿmal* (travaille !). Il est clair que le mot mal écrit a la même orthographe que les deux mots candidats à l'exception de la lettre أَ *ʾa* et إِ *ʾi*.

Correction des erreurs afin de corriger les mots avec la lettre *Alif* ا mal orthographiés, nous avons utilisé la distance de cosinus afin de mesurer la similitude entre deux vecteurs multidimensionnels. En effet, chaque mot dans les corpus d'apprentissage du modèle de langage est projeté dans un espace de 200 dimensions de sorte que les mots qui partagent le même contexte dans le corpus sont situés à proximité les uns aux autres dans l'espace. Cela est obtenu avec word2vec [Mikolov *et al.*, 2013]. Cette technique nous a permis de corriger un mot mal orthographié par un autre qui partage le même contexte.

3.6.2 Apprentissage

Après avoir normalisé les données d'apprentissage, le modèle de langage que nous proposons est la combinaison linéaire de deux modèles de langage de type n-grammes. Dans un tel modèle, la probabilité de n'importe quelle séquence de mots W est calculée selon la formule suivante :

$$P(W) = \lambda_1 P_{ML_1}(W) + \lambda_2 P_{ML_2}(W) \tag{3.1}$$

les λ_i sont des poids de pondération associés à chaque modèle de langage. Ils sont estimés sur un corpus de validation de sorte que la perplexité sur ce corpus soit minimale. La perplexité d'un modèle de langage mesure la capacité de ce modèle à prédire les mots d'un texte. Elle peut être vue comme un facteur de branchement dans un graphe où le modèle de langage joue le rôle de l'algorithme de recherche et les nœuds sont les mots à prédire. Dans ce cas, l'estimation de chaque mot revient à choisir entre PP (avec PP est la valeur de la perplexité) mots équiprobables. Il est clair que $PP \leq taille\ vocabulaire$ et plus la valeur de la perplexité est petite, plus le nombre de possibilités entre les mots équiprobables est réduit et donc meilleur sera le modèle de langage.

Le premier modèle de langage est appris sur le corpus Gigaword et le deuxième est appris sur la transcription textuelle des données acoustiques. Le nombre des n-grammes dans le modèle final est présenté dans le tableau 3.9.

n-gramme	Sans élagage	*Cutoff*	*Stolcke pruning*
1-gramme	95 589		
2-grammes	69 307 k	20 164 k	2 449 k
3-grammes	327 302 k	22 283 k	1 395 k
4-grammes	586 722 k	4 967 k	192 k

TABLE 3.9 – Nombre des n-grammes avant/après l'élagage dans le modèle de langage final.

Vu que le nombre des n-grammes dans le modèle de base (sans élagage) est très grand et à cause des contraintes de mémoire, nous avons utilisé deux techniques d'élagage pour réduire le nombre des n-grammes. Dans la première technique, un *cutoff* est appliqué pour supprimer les n-grammes peu fréquents dont la probabilité est inférieure à 10^{-9}. La deuxième technique est basée sur un élagage plus agressif, il s'agit de *Stolcke pruning* [Stolcke, 2000]. Elle est basée sur la minimisation de l'entropie relative entre le modèle initial et le modèle élagué.

3.7 Mise en œuvre

Le développement de notre système ALASR est basé sur la boîte à outils Kaldi [Povey *et al.*, 2011]. Cette boîte implémente la plupart des techniques proposées dans le cadre de la modélisation acoustique de notre système. Concernant la modélisation du langage, l'outil SRILM [Stolcke *et al.*, 2011] est utilisé pour apprendre le modèle de langage. Un aperçu de notre démarche pour la mise en œuvre de différents modèles est présenté dans ce qui suit.

3.7.1 Modèle de langage

Le processus de l'apprentissage du modèle de langage commence par la phase de normalisation des données présentées dans la section 3.6.1. Pour entraîner les différents modèles (sans et avec élagage), nous avons utilisé la boîte à outils SRILM [Stolcke *et al.*, 2011]. La plupart des modèles de langage implémentés dans SRILM sont de type n-gramme avec le support de plusieurs techniques de lissage ; il implémente aussi diverses méthodes pour interpoler et adapter les modèles de langage, ainsi que des techniques pour l'élagage. L'impact des techniques de l'élagage proposées dans le cadre de nos travaux sur les performances du système ALASR sera étudié en détail dans la section 3.8.

3.7.2 Lexique de prononciations

Le lexique de prononciations est composé de 95k mots sélectionnés à partir des données textuelles. Les prononciations de chaque mot sont celles récupérées à partir des lexiques externes (ceux proposés par [Ali *et al.*, 2014], NEMALR et NetDC). Le nombre d'entrée dans le lexique final dépend de la technique utilisée pour la modélisation des aspects acoustiques caractérisant la langue arabe (modélisation implicite et/ou explicite des voyelles courtes, longues et de la gémination). L'impact de ces aspects sur la modélisation acoustique de la langue arabe sera présenté dans la section 3.8.

3.7.3 Modèle acoustique

Les modèles acoustiques proposés sont entraîné avec Kaldi [Povey *et al.*, 2011]. Kaldi est une boîte à outils libre conçue pour traiter les données vocales. Bien qu'il ait été conçu au départ

pour les applications liées à la reconnaissance de la parole, il est utilisé dans plusieurs tâches relatives à la voix, à savoir la reconnaissance de locuteurs, la segmentation et le regroupement en locuteurs, l'identification de langage à partir du signal, etc.

Le développement du système ALASR en utilisant Kaldi est basé sur une *recette* comportant les quatre étapes suivantes :

- Préparation des données : en plus des données acoustiques brutes (la collection de fichiers audio avec leur transcription), on doit préparer des métadonnées nécessaires pour l'apprentissage du modèle acoustique et du modèle de langage. Les métadonnées acoustiques incluent des informations sur les locuteurs, la liste de tous les fichiers audio avec leur identifiant, la transcription pour chaque segment de parole et la correspondance entre les segments de la parole et les locuteurs. Les métadonnées linguistiques contiennent le lexique de prononciation, la liste de tous les phonèmes et la liste de toutes les unités acoustiques non linguistiques.
- Extraction des paramètres acoustiques : Kaldi propose plusieurs scripts pour l'extraction et la transformation des paramètres acoustiques. Nous avons utilisé dans notre cas les paramètres MFCC et fMLLR pour l'apprentissage adaptatif SAT.
- Apprentissage du modèle acoustique : plusieurs modèles GMM-HMM ont été entraînés afin de générer l'alignement entre les données et les unités acoustiques et d'initialiser l'apprentissage des modèles basés sur les réseaux de neurones DNN-HMM. Nous avons proposé deux architectures neuronales, une est basée sur le perceptron multicouche et l'autre sur les réseaux TDNN.
- Décodage : Kaldi construit un graphe pour accomplir la tâche de la reconnaissance de la parole.

3.7.4 Décodage

Les transducteurs à états finis pondérés, *weighted finite-state transducer* (WFST) sont utilisés par Kaldi pour représenter, d'une part, le modèle acoustique, le modèle de langage et le lexique, et pour établir, d'autre part, le lien entre ces modèles. Kaldi construit un énorme graphe qui s'appelle le graphe de décodage *HCLG* pour accomplir la tâche de la reconnaissance de la parole. Le graphe HCLG est construit à partir de quatre WFSTs simples représentant chacun un modèle de base comme le montre le tableau 3.10 [Mohri *et al.*, 2008].

	Automate WFST	**Séquence d'entrée**	**Séquence de sortie**
G	Le modèle de langage	Mots	Mots
L	Le lexique	Phonèmes	Mots
C	Phonèmes contextuels	Phonèmes contextuels	Phonèmes
H	HMM	États du HMM	Phonèmes contextuels

TABLE 3.10 – Les WFSTs utilisés par Kaldi pour la construction du graphe de décodage *HCLG*.

La meilleure séquence de mots $\hat{W}$ est déterminée en maximisant la combinaison de deux scores : le score du modèle acoustique $P(O|W)$ et celui du modèle de langage $P(W)$. Cependant, ces deux scores sont estimés sur des données d'apprentissage différentes, ce qui conduit à une échelle de probabilité différente. En effet, le score du modèle de langage est plus grand que celui calculé par le modèle acoustique. Pour ajuster l'échelle des deux probabilités, le score du modèle de langage est pondéré comme le montre l'équation 3.2.

$$\hat{W} = \arg\max_{W} P(O|W)P(W)^{w_{ML}} \tag{3.2}$$

Le poids w_{ML} est estimé sur le corpus de validation en le variant entre 10 et 30, dans notre cas, avec un pas de 1. En plus de ces deux scores, nous avons ajouté un autre celui de la pénalité de mots pour ne pas pénaliser les courtes phrases. En effet, si deux séquences de mots se prononcent de la manière, alors le modèle acoustique a tendance à choisir la séquence la plus longue ce qui est équivalent à choisir de courts mots. Prenons à titre d'exemple les deux séquences de mots إمعة *ā'mʿh* (suiveur) et إن مع *ā'n mʿ* (si avec), le modèle acoustique privilège la deuxième séquence qui est composée de deux mots. Du point de vu acoustique, elles se prononcent de la même manière, mais du point de vu linguistique, elles sont deux séquences de mots différentes. La pénalité de mots P_w est intégrée dans le calcul du score final selon la formule 3.3.

$$\hat{W} = \arg\max_{W}(P(O|W)P(W)^{w_{ML}}P_w^{|W|}) \tag{3.3}$$

Pour favoriser les mots longs (cela est équivalent à favoriser les séquences de mots les plus courtes), nous avons varié P_w entre 0 et 1.

3.8 Résultats et discussion

Dans cette section, nous présentons les différentes expérimentations menées pour évaluer notre système ALASR. Pour chaque composant du système, différentes variantes et/ou différents aspects sont évalués, à savoir l'architecture du modèle acoustique, les techniques de modélisation de la prononciation et les aspects relatifs à la modélisation du langage.

Parmi les architectures que nous avons proposées pour la modélisation acoustique, on trouve les modèles à base de réseau de neurones. Vu que ces modèles sont fortement impactés par les modèles GMM-HMM, notre démarche expérimentale commence par trouver la meilleure configuration qui assure le meilleur modèle acoustique. Ce dernier sera utilisé comme un modèle de base pour initialiser l'apprentissage des modèles à base de réseaux de neurones.

3.8.1 Modèles GMM-HMM

Les différents modèles GMM-HMM présentés dans la section 3.4.1 sont évalués en fonction des aspects relatifs à la modélisation de langage et à la prononciation. Selon les différentes techniques proposées, nous avons testé une série de configurations comme le montre le tableau 3.11 afin de trouver le meilleur modèle GMM-HMM.

Configurations		conf1	conf2	conf3	conf4	conf5	conf6	conf7	conf8
Modélisation de langage	Normalisation données textuelles	Oui	Oui	Oui	Oui	Oui	Oui	Oui	Oui
	Auto-correction de *hamza* ء	Non	Non	Non	Non	Non	Non	Non	Non
	Ordre du modèle de langage	4	4	3	2	2	2	2	2
	Élagage	agressif	*cutoff*	*cutoff*	*cutoff*	Non	*cutoff*	*cutoff*	*cutoff*
Modélisation de la prononciation	Modélisation explicite de voyelles courtes	Non	Non	Non	Non	Non	Oui	Oui	Oui
	Modélisation explicite de la gémination	Non	Non	Non	Non	Non	Non	Oui	Oui
	Modélisation explicite de voyelles longues	Non	Non	Non	Non	Non	Non	Non	Oui
	Nombre de mots dans le lexique	95k	95k	95k	95k	95k	95k	95k	95k
	Nombre total de variantes de prononciation	95k	95k	95k	95k	95k	475k	484k	485k

TABLE 3.11 – Différentes configurations pour tester nos approches de modélisation du langage et de la prononciation.

La première configuration représente le système de base. Les configurations *conf2* jusqu'à *conf5* ont pour objectif d'étudier l'impact des aspects relatifs à la modélisation de langage sur les performances de ALASR, en particulier l'élagage, l'ordre du modèle de langage et le processus de normalisation de données. Dans les configurations *conf6* à *conf8*, nous étudions l'impact de la modélisation explicite de voyelles longues, courtes et du *Chad-da* sur la modélisation acoustique de la langues arabe standard.

Les résultats de reconnaissance de la partie de validation de nos corpus oraux selon les différentes configurations sont présentés dans la figure 3.6.

Il est clair que le modèle acoustique *triphone3-mmi* basé sur le critère *MMI* est le meilleur modèle pour toutes les configurations. Rappelons que ce modèle est entraîné pour maximiser l'information mutuelle entre une observation acoustique O et la séquence de mots W_r qui lui correspond. Cela permet de maximiser, d'une part, la probabilité de l'hypothèse correcte, et de diminuer, d'autre part, la probabilité des autres hypothèses qui ne sont pas correctes.

Les modèles de langage d'ordre supérieur 3 et 4-grammes (*conf2* et *conf3*) assurent de meilleurs résultats de reconnaissance par rapport aux modèles d'ordre inférieur 2-grammes (*conf4* et *conf5*). Aussi, nous remarquons que l'élagage de type *cutoff* ne dégrade pas les résultats de reconnaissance (une différence absolue de 0,1% entre le modèle *triphone1* de la *conf4* et *conf5*) contrairement à l'élagage agressif *stolcke pruning* qui entraîne une dégradation importante dans le WER (une différence absolue de 3% entre le modèle *triphone1* de la *conf1* et *conf2*). Cela est justifié par le nombre des n-grammes dans chaque modèle ; le nombre des n-grammes dans le modèle élagué avec un *cutoff* représente 5% du modèle initial, contrairement au modèle élagué avec *stolcke pruning* où le nombre des n-grammes représente seulement 0,04% du modèle initial.

L'autre remarque importante concerne la modélisation explicite des voyelles lors de la modélisation acoustique (*conf6*, *conf7* et *conf8*). Bien que cette approche augmente le nombre d'entrées dans le lexique, ce qui rend le processus du décodage plus long, elle améliore considérablement la sortie du système (l'amélioration absolue du WER varie entre 14,1% et 15.9%). Cela permet au modèle acoustique de mieux aligner le signal avec les unités acoustiques en augmentant la précision de la position des voyelles dans le signal. Ainsi, la modélisation explicite de la gémination et des voyelles longues (*conf7* et *conf8*) améliore les performances du système par rapport au modèle de la *conf6*.

Il est à noter que le poids du modèle de langage w_{ML} et la pénalité de mots P_w (voir l'équation

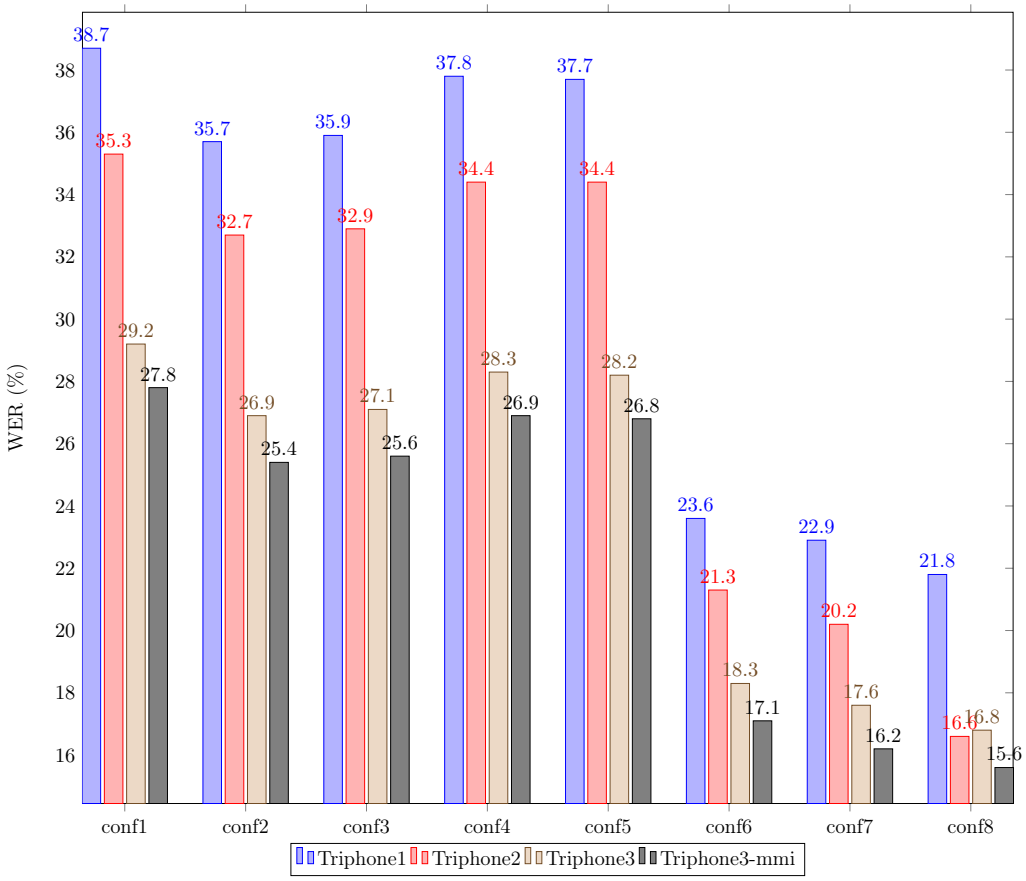


FIGURE 3.6 – Taux d'erreur obtenus sur la partie de validation selon les différentes configurations proposées.

3.3) ont été optimisés pour chaque modèle *triphone* séparément. Nous avons trouvé que le meilleur score WER est celui obtenu en fixant le coefficient de pondération w_{ML} du modèle de langage entre 14 et 17 et la pénalité de mots P_w à 0,5.

Le modèle *triphone-mmi* étant le meilleur modèle GMM-HMM (le taux d'erreur est des 15,6% avec la *conf8*), est utilisé pour initialiser l'apprentissage des modèles à base de réseaux de neurones dont les résultats sont présentés dans la section suivante.

3.8.2 Modèle à base des réseaux de neurones

Les modèles à base des réseaux de neurones sont utilisés dans notre cas pour améliorer la modélisation acoustique de l'arabe standard. Les autres modèles (modèle de langage et de la prononciation) ne changent pas et ce sont ceux de la *conf8*. Les résultats de reconnaissance en termes du WER obtenus avec les deux architectures neuronales décrites dans la section 3.4.2 sont présentés dans le tableau 3.12.

Architecture	**Fonction objective**	**WER**	**OOV**
Perceptron multicouche	Entropie-croisée	14,0 %	2,3 %
	MPE	13,5 %	
	sMBR	13,4 %	
TDNN	sMBR	11,8 %	

TABLE 3.12 – Taux d'erreur obtenus sur la partie de validation de nos corpus avec les différentes architectures neuronales.

D'après les résultats du tableau 3.12, une amélioration absolue de 1,7% est obtenue avec le modèle à base de réseaux de neurones TDNN sachant que l'intervalle de confiance est de ±0.75%. Ces résultats sont attendus vu que le modèle à base de TDNN capture mieux le contexte des observations acoustiques. Pour chaque observations O_t à l'instant t, la prédiction du triphone qui lui correspond est basée sur une fenêtre de contexte $[t-16; t+12]$ contrairement au modèle à base du perceptron multicouche où la fenêtre de contexte est fixée à $[t-5; t+5]$. Nous avons, aussi, trouvé que le meilleur moyen pour estimer les paramètres du modèle est d'utiliser une fonction objective qui minimise la distance entre les triphones de la séquence de référence W_r et la prédiction du modèle W (le critère sMBR).

En analysant la sortie du système de reconnaissance basé sur le modèle TDNN et sur les modèles de la *conf8*, nous avons trouvé que 33% des substitutions concernent les mots avec le symbole *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا. Les mots avec le symbole *hamza* ء sont souvent remplacés avec d'autres mots sans *hamza* ء; nous donnons dans le tableau 3.13 quelques exemples de substitution de mots.

Mot	**Remplacé par**	**Fréquence**	**Traduction en français**
أيضَا *ʾyḍā*	ايضَا *āyḍā*	12	Aussi
إسرَائِيل *ʾisrāʾiyl*	اسرَائِيل *āsrāʾiyl*	19	Israël
أن *ʾn*	ان *ān*	43	Que

TABLE 3.13 – Quelques exemples de substitution de mots avec le symbole *hamza* ء.

En corrigeant l'écriture du symbole *hamza* ء dans la sortie du système de reconnaissance en utilisant le processus proposé dans la section 3.6, nous obtenons les résultats présentés dans le tableau 3.14.

Modèle	**Validation (WER)**	**Test (WER)**
TDNN+conf8	11,8	12,7
Correction	11,3	12,5
Rescoring	11,4	12,4
Rescoring+correction	**10,6**	**12**

TABLE 3.14 – Taux d'erreur avant et après la correction de l'écriture de *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا.

La ligne *rescoring* dans le tableau 3.14 signifie que le score du modèle de langage est recalculé avec un modèle d'ordre supérieur. Cette procédure consiste à remplacer le score du modèle de

langage 2-grammes dans le graphe de décodage par le score du modèle 4-grammes. Avec cette approche, nous visons à améliorer la probabilité des séquences de mots ce qui améliore, par conséquent, la sortie de notre système. Le *rescoring* et la correction de l'écriture de *hamza* ء conduisent à une amélioration absolue de 1,2% sur la partie de validation et de 0,7% sur la partie du test.

Comparer notre système avec les autres systèmes état de l'art est une tâche complexe, car les performances d'un système dépendent des conditions dans lesquelles il est évalué. Nous présentons dans le tableau 3.15 les résultats publiés pour quelques systèmes état de l'art de reconnaissance automatique de la parole arabe développés dans les cadre du projet Gale.

Système	**Modèle**	**Taille vocabulaire**	**Données acoustiques**	**Données textuelles**	**WER**
[Soltau *et al.*, 2014]	Sans voyelles	589 k	135 h+1000 h (non supervisé)	56 M	17,0 %
[Soltau *et al.*, 2014]	Avec voyelles	589 k	135 h+1000 h (non supervisé)	56 M	14,0 %
[Ali *et al.*, 2014]	Avec voyelles	526 k	194 h	1,4 M	15,8 %
[Smit *et al.*, 2018]	Sans voyelles	900 k	1200 h	1,5 M	13,2 %
[Khurana et Ali, 2016]	Sans voyelles	900 k	1200 h	138 M	14,7 %
ALASR [Menacer *et al.*, 2017b, Menacer *et al.*, 2017c]	Avec voyelles	95 k	52 h	1 000 M	12,7 %

TABLE 3.15 – Résultats état de l'art de reconnaissance automatique de la parole arabe.

Les performances des systèmes de reconnaissance automatique de la parole sont étroitement liées aux ressources utilisées. On peut remarquer que dans la plupart des systèmes, des milliers d'heures sont utilisées pour l'apprentissage du modèle acoustique, tandis que la taille des ressources textuelles est dans l'ordre de millions de mots. Ainsi, le WER pour les systèmes de reconnaissance automatique de la parole actuels est dans l'ordre de 15% qui est très comparable aux 10% du WER pour les systèmes anglais. Cela prouve ainsi que notre système ALASR assure des résultats comparables à ceux des systèmes de l'état de l'art.

3.9 Conclusion et discussion

Dans ce chapitre, nous avons présenté notre première contribution dans le cadre de la thèse, il s'agit du développement d'un système de reconnaissance automatique de la parole pour l'arabe standard que nous avons appelé ALASR. Ce système est basé sur les approches statistiques qui sont généralement indépendantes de la langue et qui sont tout à fait applicables pour l'arabe standard. Cependant, la langue arabe présente quelques caractéristiques que nous devons prendre en considération afin de booster les performances du système de reconnaissance.

Le premier défi que nous avons relevé, dans le cadre de la modélisation acoustique, était l'absence de l'indication des voyelles dans le texte écrit. L'arabe est une langue riche en consonnes et pauvre en voyelles. Parmi les 34 phonèmes, seulement 3 sont des voyelles dont chacune possède une forme longue et courte. L'indication des voyelles est restreinte aux textes religieux, poétiques et aux ouvrages destinés aux enfants pour faciliter la prononciation des mots. La plupart des textes disponibles sont écrits les diacritiques (qui servent à indiquer les voyelles courtes). L'absence de ces dernières rend difficile leur apprentissage par le modèle acoustique car il est difficile de savoir quelles voyelles sont utilisées, et à quelle position dans le mot. Nous avons montré que la

modélisation explicite des voyelles est la meilleure approche pour apprendre le modèle acoustique pour l'arabe standard.
Nous avons, aussi, étudié le phénomène de la gémination dans le cadre de la modélisation acoustique. La gémination, ou *chad-da* en arabe, consiste en un redoublement de consonne. Dans le cas où un locuteur parle lentement, les consonnes géminées possèdent une durée accrue perceptible à l'oreille. Nous avons trouvé que la modélisation explicite de *chad-da* dans le lexique de prononciations améliore le modèle acoustique. Le même travail a été réalisé pour les voyelles longues et nous avons trouvé que leur modélisation assure de meilleurs résultats.

Bien que l'arabe standard soit soumis à des règles linguistiques bien définies, il arrive des cas où l'écriture de certains mots est simplifiée en remplaçant une lettre par une autre ou en omettant le symbole *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا. Cela a un impact direct sur le modèle de langage. Nous avons proposé une phase de normalisation de données textuelles pour harmoniser l'orthographe de nos corpus. Elle consiste à corriger les erreurs d'orthographe ainsi que les variations d'écriture se rapportant au *hamza* ء .

La présence de plusieurs variantes parlées de la langue arabe est un autre défi que nous devons surmonter. L'arabe standard n'étant pas la langue maternelle des arabes, son utilisation est souvent restreinte à l'enseignement dans les écoles, les livres, les journaux, les magazines et les médias officiels. En l'occurrence, les arabes utilisent le dialecte dans leurs conversations quotidiennes. Le dialecte est une forme informelle, intermédiaire entre la langue standard et le patois et chaque région du monde arabe a sa propre variante. Dans le chapitre suivant, nous présentons notre travail pour adapter le système ALASR pour une variante du dialecte : le dialecte algérien.

Chapitre 4

L'adaptation d'ALASR pour le dialecte Algérien

Sommaire

Comme on l'a vu dans le chapitre précédent, l'arabe standard moderne n'est pas la langue maternelle dans le monde arabe. On utilise plutôt dans les conversations quotidiennes une autre variante de langue connue sous le nom de dialecte.

Ces dernières années et avec l'émergence des sites communautaires, les chercheurs travaillant sur le traitement automatique des langues s'intéressent de plus en plus au traitement automatique des dialectes arabes puisque c'est la forme la plus utilisée dans les réseaux sociaux et les émissions télévisées.

Dans ce chapitre, nous nous intéressons à l'adaptation de notre système ALASR au dialecte algérien. Ce dernier est l'un des dialectes maghrébins qui se caractérisent par leur difficulté

à être traités par les systèmes de reconnaissance automatique de la parole. Du point de vue linguistique, cette difficulté est principalement due au vocabulaire du dialecte algérien qui est très évolutif et influencé par d'autres langues, notamment par la langue française. Du point de vue computationnel, le manque de ressources nécessaires pour apprendre les différents modèles est le problème majeur qui rend difficile la modélisation du dialecte algérien. Notre approche pour remédier à ce problème consiste à tirer profit de données provenant d'autres langues qui ont un impact sur le dialecte algérien [Menacer *et al.*, 2017c]. Nous présentons, dans ce chapitre, nos approches pour intégrer et utiliser ces données dans la modélisation linguistique et acoustique du dialecte algérien.

4.1 Le dialecte algérien

Il existe deux langues officielles en Algérie, à savoir l'arabe standard moderne et le tamazight. Bien que l'arabe standard soit compréhensible par la majorité de la population algérienne parce qu'elle est enseignée dans les écoles, elle n'est pas la langue maternelle dans ce pays. On utilise principalement le tamazight et le dialecte dans les conversations quotidiennes. En outre, le français, bien qu'il n'ait pas de statut officiel, est largement utilisé par l'administration, la culture, les médias et l'éducation (dès l'école primaire). Pendant très longtemps, l'utilisation du dialecte était restreinte aux discussions informelles dans les familles, voisins et amis. Ces dernières années, avec l'émergence des réseaux sociaux et la libéralisation de l'audiovisuel en Algérie, désormais on utilise le dialecte dans les médias et plus particulièrement dans les réseaux sociaux.
Géographiquement, l'Algérie est connue pour sa grande superficie d'où la diversité des dialectes selon les régions ; on cite alors le dialecte qu'on a sélectionné pour nos travaux : le dialecte algérois reconnu pour être le dialecte de la régions du centre.

À l'instar des autres dialectes arabes, les mots empruntés d'autres langues tiennent une place importante dans le vocabulaire du dialecte algérien. Parmi les langues qui influencent le dialecte algérien on trouve le français, le berbère, le turque, l'espagnole et l'italien. Les emprunts de la la langue française sont majoritaires par rapport aux emprunts d'autres langues. Cela est justifié par les termes assez anciens empruntés durant la période de la colonisation française (1830-1962), et ceux plus modernes qui ne cessent d'être introduits dans le dialecte algérien [Guella, 2011]. Des exemples de quelques emprunts lexicaux sont présentés dans le tableau 4.1.

Emprunts	**Origine**	**Traduction en arabe**	**Traduction en français**
فَكرون *fakrwn*	berbère	سلحفَاة *slḥfāh*	tortue
سعَايجي *sʿāyğy*	turque	سَاعَاتي *sāʿāty*	horloger
ميزرية *myzryh*	espagnol	بُؤس *buʾs*	misère
كَازِرنة *kaāzirnh*	français	ثكنة *ṯknh*	caserne

Table 4.1 – Exemples d'emprunts lexicaux dans le dialecte algérien.

Ces mots ont été facilement et naturellement incorporés dans le dialecte algérien en respectant la structure morphologique de la langue arabe standard. Ce qui a conduit à l'altération de ces mots au niveau lexical et phonologique. Pour les mots d'origine arabe qui constituent la majorité des mots du dialecte algérien, c'est surtout leur prononciation qui est altérée. Toutes ces caractéristiques font que le dialecte algérien figure parmi les dialectes les plus difficiles à modéliser pour les systèmes de reconnaissance de la parole.

4.2 Défis pour les systèmes de reconnaissance automatique de la parole pour le dialecte

Développer un système de reconnaissance automatique de la parole pour n'importe quelle langue nécessite une grande quantité de données acoustiques et textuelles. Malheureusement ces données sont peu disponibles pour le dialecte algérien, ce qui permet de le classer dans les langues peu dotées en ressources.

Le traitement automatique des langues peu dotées en ressources est loin d'être facile, du fait qu'il nécessite des techniques plus sophistiquées qui vont bien au-delà de l'application des *recettes* habituellement utilisés pour la modélisation des langues bien dotées en ressources. Avant de présenter les techniques d'apprentissage utilisées pour développer un système de reconnaissance automatique de la parole pour le dialecte, nous discutons, dans les sections suivantes, les aspects acoustiques et linguistiques qui caractérisent le dialecte algérien et que nous devons prendre en considération avant d'adapter dans notre système ALASR au dialecte algérien.

4.2.1 Aspects acoustiques

Dans le dialecte algérien, les mots empruntés utilisent des sons qui n'existent pas en arabe standard et qui sont souvent utilisés en français. Le système phonétique du dialecte est composé, en plus des 28 consones et des 6 voyelles de l'arabe standard, de 4 consonnes et de 9 voyelles qu'on peut les trouver dans la langue française (voir tableau 4.2).

Phonèmes	**Exemple**	**Prononciation**	**Traduction en français**
/E/	مَاستَار *māstār*	/m a s t E r/	master
/e/	مَارشي *māršy*	/m a r S e/	marché
/g/	بَاقَاج *bāvāǧ*	/b a g a Z/	bagage
/p/	پورتَابل *pwrtābl*	/p u r t a b l/	portable
/v/	پوفوَار *pwfwār*	/p u v w a r/	pouvoir
/y/	سكوريتي *skwryty*	/s e k y r i t e/	sécurité
/a∼/	سورمون *swrmwn*	/s y r m a∼/	sûrement
/2/ /9/ /@/	شومور *šwmwr*	/S u m 2 r/	chômeur
/E/	لكوزَا *lkwzā*	/l k u z E/	le cousin
/o∼/	كونتر *kwntr*	/k o∼ n t r/	contre
/Z/	جيست *ǧyst*	/Z y s t/	juste

Table 4.2 – Phonèmes français utilisés en dialecte algérien avec un exemple de mot dialectal pour chaque phonème.

Comme on l'a mentionné dans la section précédente, la prononciation des mots dialectaux, que ce soit des mots empruntés ou des mots arabes, peut être la même que celle de la langue originale ou elle peut être altérée. C'est pourquoi, on peut classer ces mots en deux catégories selon leur prononciation par rapport à la langue originale. Dans le tableau 4.3, nous donnons quelques exemples de mots dialectaux où la prononciation est la même que la langue originale (les deux premières lignes) et des mots où la prononciation est altérée (les deux dernières lignes).

Mot dialectal	**Traduction**	**Prononciation (dialecte)**	**Prononciation (langue originale)**
تيكي *tyky*	Ticket	/t i k e/	/t i k e/
النَاس *ālnās*	Les gens	/? a l n n a s/	/? a l n n a s/
قَاطو *vāṭw*	Gâteau	/g a t‘ u/	/g a t o/
بيت *byt*	Maison	/b i t/	/b a y t/

TABLE 4.3 – Exemples de quelques mots en français utilisés dans le dialecte algérien.

Le problème qui se pose concerne la modélisation acoustique des phonèmes du dialecte algérien sachant qu'il n'existe aucun corpus oral transcrit de dialecte pour permettre l'apprentissage direct par des *recettes* standard ; et comment tirer profit du fait que l'on retrouve les phonèmes du dialecte algérien soit dans la parole arabe, soit dans la parole française.

4.2.2 Aspects linguistiques

Bien que le dialecte algérien ait une morphologie moins complexe que l'arabe standard, il présente des défis plus exigeants que nous devons surmonter. En effet, le dialecte algérien est principalement parlé, il n'existe pas de normes ni de règles pour l'écrire. Ces dernières années, avec l'apparition des réseau sociaux, les algériens ont tendance à écrire les mots dialectaux tel qu'ils sont prononcés sans accorder la moindre attention aux règles linguistiques de l'arabe standard puisqu'il n'existe pas de règles pour le dialecte. De ce fait, il arrive des cas où un mot peut avoir plusieurs formes d'écriture. Citons à titre d'exemple le mot مَعلَابَاليش *maʿlābālyš* (je ne sais pas) qui peut être écrit مَاعلَابَاليش *māʿlābālyš* معلَابَاليش *mʿlābālyš* مَاعلبليش *māʿlblyš* مَاعلَابَالش *māʿlābālš* ainsi que d'autre formes dépendantes de la personne qui écrit (son accent, sa localisation géographique, son niveau intellectuel, etc.). Ce phénomène est considéré normal en dialecte, bien qu'il soit considéré comme une erreur lexicale dans l'arabe standard.

Le dialecte algérien étant riche de mots empruntés d'autres langues, essentiellement celles basées sur le script latin, l'alphabet latin vient remplacer l'alphabet arabe dans les messages pour faciliter la compréhension et l'écriture de ces derniers. Avec l'utilisation du script latin, certains mots arabes deviennent difficile à cause de la présence de consonnes non existantes en latin ; les internautes ont trouvé une alternative pour ce problème en combinant plusieurs lettres de l'alphabet latin ou bien en changeant certaines combinaisons en chiffres. Parmi les caractères remplacés par des chiffres on cite : ح remplacé par 7, ع par 3, ط par 6 et ق par 9. Ainsi, le mot مَعلَابَاليش *maʿlābālyš* s'écrit en script latin *ma3labalich*.

4.3 Données utilisées

Le principal défi auquel nous sommes confrontés est le manque des données textuelles et orales pour le dialecte algérien. Le moyen le plus simple pour collecter les données textuelles est d'exploiter le contenu des sites communautaires. Cependant, les corpus oraux avec leur transcription sont plus difficiles à collecter pour le dialecte.

4.3.1 Données textuelles

Le dialecte algérien est principalement parlé, et n'a aucune règle d'écriture conventionnelle. Une façon pour collecter les données textuelles est de les récupérer à partir des réseaux sociaux. Deux corpus contenant du dialecte algérien ont été récemment constitués : le corpus PADIC [Meftouh *et al.*, 2015, Meftouh *et al.*, 2018] et CALYOU [Abidi et Smaïli, 2017, Abidi *et al.*, 2017].

PADIC est une collection de 6 400 phrases en arabe standard avec leurs traductions dans plusieurs dialectes arabes (le dialecte algérois l'objet de notre travail, un dialecte du nord-est algérien, un tunisien, un marocain, un palestinien et un syrien). Ce corpus a été développé manuellement en traduisant des phrases de conversations en dialecte algérois vers l'arabe standard et vers les autres dialectes.

CALYOU est un corpus de taille plus importante que le précédent, contenant des commentaires de vidéos algériennes sur YouTube. Ces commentaires ont été collectées en se basant sur une liste de mots-clés correspondant à des évènements ou à des personnalités connus principalement par les algériens. Le corpus résultant contient plus d'un million de commentaires.

Des statistiques sur les deux corpus sont données dans le tableau 4.4

Corpus	Nombre de phrases	Nombre de mots	Nombre de Mots uniques
PADIC	6,4 k	25 k	6,6 k
CALYOU	1 200 k	18 300 k	520 k

TABLE 4.4 – Statistiques sur les corpus textuels dialectaux.

4.3.2 Donnée orales

À notre connaissances, il n'existe pas de corpus oraux pour le dialecte algérien. Pour cette raison, nous avons décidé d'en construire un. Pour ce faire, notre approche consiste à sélectionner un certain nombre de phrases à partir de corpus textuels et de les enregistrer par des locuteurs natifs. Vu que cette tâche est coûteuse en temps, le corpus construit est de petite taille, il sera utilisé pour adapter les modèles déjà appris dans le système ALASR.

La procédure de construction du corpus est la suivante : nous avons sélectionné 4,6k phrases à partir de PADIC et CALYOU. Le nombre de mots dans les phrases sélectionnées varie entre 3 et 20 mots. Nous avons demandé à 7 locuteurs natifs de lire les phrases pour les enregistrer avec un microphone professionnel unidirectionnel dans une pièce silencieuse. Parmi les sept locuteurs, deux sont des femmes et leur âge varie entre 26 et 52 ans. Le corpus résultant, nommé ADIC pour *Algerian DIalect Corpus*, contient 6 heures de parole réparties en trois sous ensembles comme le montre le tableau 4.5. Il est à noter que les locuteurs de l'ensemble de test n'ont pas participé à l'enregistrement de l'ensemble d'apprentissage et de validation.

Partie	Durée	Nombre de locuteurs		
		Femme	Homme	Total
Apprentissage	240 min	1	3	4
Validation	40 min	1	1	
Test	75 min	1	2	3

TABLE 4.5 – Statistiques sur le corpus oral ADIC.

4.4 Modélisation acoustique

Le modèle à base de réseaux de neurones TDNN étant le modèle le plus performant pour l'arabe standard, nous l'avons utilisé pour la modélisation acoustique du dialecte. L'apprentissage de tels modèles nécessite une grande quantité de données acoustiques, quantité qui n'est pas disponible pour le dialecte. Pour remédier à ce problème et en partant de l'idée que l'ensemble des phonèmes du dialecte algérien fait parti de de l'ensemble des phonèmes de l'arabe standard et celui du français, nous avons décidé de tirer profit des données acoustiques de ces deux langues. Cette idée est inspirée du fait que le cerveau humain peut exploiter des connaissances déjà acquises afin de perfectionner l'apprentissage de nouvelles connaissances. Pour projeter ce principe de fonctionnement sur les réseaux de neurones, nous transférons les connaissances apprises par le modèle acoustique du système ALASR au modèle acoustique du dialecte algérien. Pour ce faire, nous proposons trois approches différentes selon la manière dont les données acoustiques de l'arabe standard et du français sont intégrées dans le processus d'apprentissage du modèle acoustique du dialecte algérien. Ces approches sont : l'apprentissage multilingue, l'apprentissage multitâche et le transfert de connaissances (voir figure 4.1).

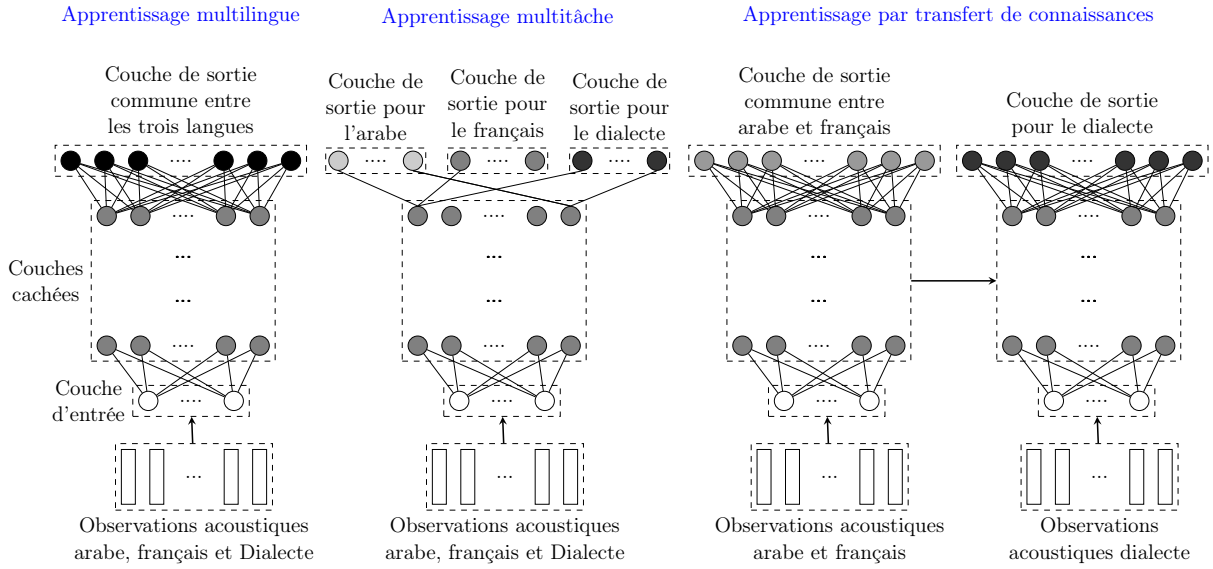


FIGURE 4.1 – Techniques d'apprentissage proposées pour le modèle acoustique.

4.4.1 Apprentissage multilingue

L'apprentissage multilingue consiste à apprendre un seul modèle acoustique à partir d'un mélange de données de plusieurs langues : le dialecte algérien, l'arabe standard et le français [12]. Le modèle acoustique final modélise, en plus des phonèmes du dialecte, 34 phonèmes de l'arabe standard et 13 phonèmes du français, les phonèmes des deux autres langues : l'arabe standard et le français. Concrètement, toutes les couches du réseau de neurones y compris celle de sortie (qui prédit la probabilité d'associer chaque trame acoustique à une unité acoustique) sont partagées entre les trois langues. Le problème qui se pose à ce niveau est le choix des unités acoustiques

12. Les données de l'arabe standard sont celles utilisées dans le système ALASR. Les donnée du français ont été extraites à partir du corpus ESTER [Galliano *et al.*, 2009], et les données dialectales font partie de notre corpus ADIC.

qui vont être prédites par la couche de sortie sachant que la langue arabe et la langue française partagent quelques sons entre elles (voir tableau 4.6). Les phonèmes en gras dans le tableau 4.6 sont les phonèmes en commun entre l'arabe standard et le français, il existe au total 14 consones et 3 voyelles communs aux deux langues. Les phonèmes en bleu sont ceux utilisés dans le dialecte algérien. Et les phonèmes en noir sont ceux de la langue française qui ne sont pas utilisés en dialecte algérien.

Consones arabes	/?/	/?'/	/d'/	/D/	/D'/	/h/	/q/	/r/	/s'/	/t'/	/T/	/x/	/X/	/Z/
	/b/	**/d/**	**/f/**	**/j/**	**/k/**	**/l/**	**/m/**	**/n/**	**/s/**	**/S/**	**/t/**	**/w/**	**/z/**	**/G/**
Consones françaises	**/b/**	**/d/**	**/f/**	**/j/**	**/k/**	**/l/**	**/m/**	**/n/**	**/s/**	**/S/**	**/t/**	**/w/**	**/z/**	**/R/**
	/p/	/g/	/v/	/Z/	/J/	/N/	/H/							
Voyelles arabes	**/a/**	**/u/**	**/i/**	/a :/	/u :/	/i :/								
Voyelles françaises	**/a/**	**/u/**	**/i/**											
	/e/	/E/	/a∼/	/y/	/2/	/9/	/@/	/o∼/	/o/	/e∼/	/9∼/			

TABLE 4.6 – Système phonétique des trois langues : le dialecte algérien, l'arabe standard et le français. Phonèmes en bleu : phonèmes du dialecte ; phonèmes en gras : phonèmes en commun entre l'arabe standard et le français ; phonèmes en noir : phonèmes exclusivement français.

Nous proposons deux approches pour produire la liste des unités acoustiques à apprendre par le modèle acoustique :

Union des phonèmes : Dans cette configuration, nous prenons l'union des listes de phonèmes de l'arabe standard et du français. Le nombre d'unités acoustique à apprendre dans ce cas est de 65 phonèmes (34 phonèmes arabes et 31 phonèmes français). On peut remarquer que le nombre de phonèmes français dans le tableau 4.6 est de 35 phonèmes mais nous n'avons gardé seulement que 31 phonèmes. En effet, nous avons décidé de regrouper les phonèmes qui ont des prononciations similaires et de les modéliser par une seule unité acoustique, il s'agit de : /e/ /E/ comme dans *ses* et *seize*, /a∼/ /o∼/ comme dans *vent* et *bon* et /2/ /9/ /@/ comme dans *deux*, *neuf* et *justement*. Ce choix est justifié par le fait que les algériens ne font pas la différence entre la prononciation de ces voyelles.

Phonèmes en commun : Dans cette configuration, nous gardons qu'une seule instance pour les phonèmes communs à arabe standard et au français. La liste des unités acoustiques à apprendre dans ce cas correspond à 48 phonèmes (17 phonèmes sont communs entre les deux langues).

En intégrant des données de parole de l'arabe standard et du français dans le processus d'apprentissage du modèle acoustique pour le dialecte algérien, on risque de donner plus d'importance à ces deux langues qu'au dialecte. Pour cette raison, nous avons analysé l'impact de la quantité de données acoustiques de chaque langue utilisée dans le processus d'apprentissage. Pour ce faire, le corpus oral du dialecte est étendu en ajoutant progressivement 4 heures de chaque langue jusqu'à arriver à 44 heures d'arabe standard et 40 heures de français. La quantité optimale de données acoustiques de chaque langue est déterminée en minimisant le taux d'erreur mots sur l'ensemble de validation d'ADIC. Les paramètres du réseau de neurones sont estimés de la même manière que le modèle acoustique du système ALASR, c'est-à-dire avec l'algorithme de la descente de gradient stochastique sur des *mini-batch* de 128 et un learning rate de 0,0002.

4.4.2 Apprentissage multitâche

L'apprentissage multitâche est le fait d'entraîner le réseau de neurones sur plusieurs tâches en même temps. Cela permet au modèle acoustique d'apprendre les similarités entre différentes

langues, ce qui va lui permettre d'apprendre une langue particulière (ou les trois en même temps) d'une manière plus performante. Dans notre cas, trois tâches de reconnaissance de la parole sont apprises par un seul modèle, à savoir la reconnaissance du dialecte, de l'arabe standard et du français. La différence entre cette approche et l'apprentissage multilingue est que les unités acoustiques de chaque langue sont modélisées séparément, il n'y aucun partage au niveau de la dernière couche de prédiction. Seulement les couches cachées sont partagées entre les trois langues. L'autre différence réside dans la manière dont les paramètres sont estimés ; dans l'apprentissage multilingue, les paramètres sont estimés sur un mélange de données orales sans aucune distinction entre les trois langues. Dans l'apprentissage multitâche, les paramètres sont estimés selon deux approches :

Poids de pondération : Notre objectif est d'apprendre un modèle acoustique pour le dialecte algérien. Cela signifie que la tâche de reconnaître le dialecte algérien est plus importante que les deux autres tâches, à savoir la reconnaissance de l'arabe standard et du français. Cette approche est basée sur le travail de [Sahraeian et Compernolle, 2016]. L'idée est d'apprendre trois modèles en parallèle, un pour chaque langue. Les paramètres de chaque modèle sont ceux des couches cachées $W^{(l)}_{cachees}$ et ceux de la couche de sortie $W^{(l)}_{sortie}$. Les paramètres des couches cachées de chaque modèle sont utilisés pour ajuster les paramètres des couches cachées du modèle final $W_{cachees}$ après avoir entraîné les trois modèles sur un certain *mini-batch* (400k trames). L'ajustement de ces paramètres est fait selon la formule 4.1.

$$W_{cachees} = \sum_{l=1}^{3} p_l W^{(l)}_{cachees} \tag{4.1}$$

avec p_l est le poids associé à chaque langue l de sorte que $\sum_{l=1}^{3} p_l = 1$. Ces poids sont utilisés pour donner plus d'importance à une tâche par rapport aux autres. Les paramètres des couches cachées résultant sont redistribués comme point de départ pour la prochaine itération d'apprentissage.

Mini-batch : Dans cette approche, toutes les tâches ont le même poids. L'estimation des paramètres des couches cachées du modèle final est réalisée sur des mini-batchs contenant un mélange de 1000 trames de chaque langue.

4.4.3 Transfert de connaissances

Les couches cachées du réseau de neurones encapsulent des informations plus complexes sur les caractéristiques de l'entrée, tandis que la couche de sortie est une couche spécifique à la tâche à apprendre. Elle prédit la sortie finale à partir les informations capturées par les couches cachées. Ces informations peuvent être réutilisées pour apprendre une nouvelle tâche, c'est l'idée de l'apprentissage basé sur le transfert de connaissances. Dans le cas où une petite quantité de données est disponible, on peut apprendre le réseau de neurones sur une tâche avec une grande quantité de données et adapter le modèle obtenu pour une nouvelle tâche où peu de données d'apprentissage sont disponibles.

En pratique, cette idée revient à apprendre un modèle acoustique sur les données de l'arabe standard et/ou du français, à supprimer la dernière couche de sortie et à garder seulement les n premières couches cachées du modèle obtenu. Une nouvelle couche spécifique au dialecte algérien est ajoutée au-dessus des couches cachées du modèle initial. Cette couche prédit la probabilité d'associer chaque trame à une unité acoustique du dialecte.

Les paramètres du modèle adapté sont estimés sur les données acoustiques du dialecte algérien. Vu que les paramètres des couches cachées ont été déjà estimés sur les données de l'arabe standard et/ou du français, nous avons utilisé un *learning rate* de faible valeur (0,00002) pour les ajuster. Pour estimer les paramètres de la dernière couche spécifique au dialecte, nous avons utilisé un autre *learning rate* plus grand (0,0002). Nous avons opté pour cette approche pour ne pas trop modifier les paramètres des couches cachées en les ajustant sur les données dialectales.

4.5 Modélisation du langage

L'expression des personnes en dialecte algérien sur les réseaux sociaux est libre, ce qui conduit à des textes avec plusieurs possibilités d'écriture. Ajoutons à cela le fait que le dialecte n'obéit à aucune règle linguistique. Tous ces facteurs rendent difficile le développement d'une méthode permettant d'identifier les variantes d'écriture et de les normaliser par la suite.

4.5.1 Normalisation des données

Notre approche pour harmoniser l'orthographe des mots en dialecte consiste à trouver les variantes d'écriture de chaque mot dans nos corpus textuels et de remplacer ces formes par la forme la plus fréquente. Cette idée est surtout appliquée pour normaliser le corpus CALYOU car c'est un corpus collecté à partir de YouTube. Le corpus PADIC a été construit manuellement en fixant un certain nombre de règles d'écriture pour les mots dialectaux. Le système d'écriture dans PADIC est basé sur les lettres arabes complétées par les trois caractères (پ/p/, ڤ/v/, ڤ/g/) pour les sons non arabes. En outre, si un mot en dialecte existe en arabe standard alors il est écrit tel quel sinon il est écrit comme il est prononcé.

Pour mettre en place notre approche de normalisation de l'orthographe des mots dans le corpus CALYOU, nous avons commencé par garder seulement les commentaires de CALYOU écrits en script arabe. Tous les commentaires où au moins un mot est écrit en script latin sont supprimés. Cela réduit le nombre de commentaires de 1,2M à 612k. Par la suite, la normalisation de l'écriture des mots dialectaux est faite en se basant sur le lexique de [Abidi et Smaïli, 2018]. Ce lexique contient une liste de mots dialectaux qui sont écrits avec les deux scripts arabe et latin. Pour chaque mot, on trouve la liste de ces possibilités d'écriture dans un script différent du script initial. Un exemple de quelques entrées de ce lexique est donné dans le tableau 4.7.

Entrée	**Écritures possibles**
خويَا	khuya, khoya, 5oya, 5ouya, khouya, khoyà, khuya
mister	ميستر, مستر

TABLE 4.7 – Exemples de quelques entrées du lexique proposé par [Abidi et Smaïli, 2018]

En gardant seulement les entrées en script arabe avec leurs possibilités d'écriture et en éliminant les mots en script latin, nous avons pu extraire une liste de 200 entrées. Chaque entrée de cette liste est un groupe de mots représentant les possibilités d'écriture d'un mot. Chaque groupe est remplacé, par la suite, par la forme la plus fréquente dans le corpus CALYOU. Nous donnons dans le tableau 4.8 quelques exemples de notre liste extraite à partir du lexique de [Abidi et Smaïli, 2018].

Avec cette approche, on réduit la taille du vocabulaire en minimisant la diversité et les variantes d'écriture des mots. Ce qui améliore, par conséquent, les probabilités des séquences de

Groupe de mots	Forme plus fréquente	Traduction
فيلم فليم فلم	فيلم	Film
منفقين منَافقين	منَافقين	Hypocrites
خَاَاوة خوة خوَا خَاوَى خَاوة خَاوَا	خَاوة	Frères

TABLE 4.8 – Exemples de groupes de mots ayant plusieurs formes d'écriture dans le dialecte algérien.

mots estimées par le modèle de langage.

4.5.2 Apprentissage

L'apprentissage du modèle de langage n'est pas restreint aux données du dialectes algérien. De manière similaire à ce qui a été fait pour le modèle acoustique, nous tirons profit des données disponibles dans d'autres langues, à savoir l'arabe standard. Pour ce faire, nous interpolons linéairement quatre modèles de langage :

— Deux modèles de langage entraînés respectivement sur PADIC et CALYOU (données dialectales).
— Deux autres entraînés respectivement sur la version arabe de Gigaword et sur la transcription des données d'apprentissage de l'arabe standard.

Les coefficients d'interpolation sont estimés sur un corpus de validation composé d'un mélange de données dialectales et d'arabe standard. Les coefficients obtenus sont les suivants : 0,48 pour CALYOU, 0,22 pour Gigaword, 0,11 pour PADIC et 0,19 pour la transcription des données d'apprentissage de l'arabe standard.

L'ordre du modèle final est 2-grammes, et un élagage de type *cutoff* a été appliqué pour garder seulement les n-grammes avec une probabilité supérieure ou égale à 10^{-9}. Le nombre des n-grammes avant et après l'élagage est présenté dans le tableau 4.9

n-gramme	Sans élagage	*Cutoff*
1-gramme	125k	
2-grammes	71 825k	12 830k

TABLE 4.9 – Nombre des n-grammes avant/après l'élagage dans le modèle de langage final.

4.6 Modélisation de la prononciation

Nous avons trouvé, lors du développement du système ALASR, que la modélisation explicite des voyelles dans les variantes de prononciations des mots assurait de meilleurs résultats. Nous avons décidé de procéder de la même manière pour le dialecte algérien. Deux problèmes se posent alors :

— Comment restituer les voyelles dans le texte du dialecte ?
— Et comment générer la prononciation à partir du texte voyellé ?

Pour répondre à ces deux questions, nous avons adapté pour nos besoins une approche G2P (*Graphem-to-Phonem*) proposée par [Harrat *et al.*, 2014]. Cette approche est basée sur deux processus :

— Le restitution des voyelles dans le texte écrit. Ce problème est vu comme un problème de traduction automatique où la langue source est un ensemble de phrases non voyellées et

la langue cible est un ensemble de phrases avec voyelles. Un système de traduction automatique statistique a été entraîné sur un corpus parallèle de phrases en dialecte algérien non diacritisées et diacritisées. Comme ce corpus parallèle a été construit manuellement et que la tâche de voyellisation est coûteuse en temps, ce corpus contient seulement 4k phrases. La précision obtenue avec cette approche est de 98% au niveau de caractères et de 96% au niveau de mots.

— Une fois les diacritiques restituées, des règles sont utilisées pour produire la prononciation de mots dialectaux. En plus des règles proposées dans [Harrat *et al.*, 2014], nous avons dû ajouter des règles pour traiter quelques cas particuliers. En effet, dans le cas où le système de traduction échoue à diacritiser un mot, alors nous avons ajouté quelques exceptions pour générer la prononciation à partir du mot non-diacritisé. Ces cas sont listés dans le tableau 4.10. La première entrée du tableau 4.10 signifie que le phonème /j/ est remplacé par un /i/ s'il est entre deux consonnes.

Condition d'application	**Remplacer par**	**Exemple**	**Traduction**
(C) j (C)	(C) i (C)	تحيبيلنَا /t Z i b i l n a/	tu nous ramènes
(C) ? (C)	(C) a: (C)	نَاوي /n a: w i: /	il a l'intention
(C) w (C)	(C) u (C)	پوفوَر /p u v w a r/	pouvoir
(C)(?|w|j)$	(C) (a|u|i)$	تحبي /t X b i/	tu aimes

TABLE 4.10 – Règles ajoutés dans notre modélisation des prononciations. C désigne une consonne et $ désigne une fin de mot.

Le lexique du dialecte algérien est composé de l'union du lexique du système ALASR (95k mots et 485K prononciations), avec tous les mots du corpus PADIC (6,6k mots) et les 50k mots les plus fréquents du corpus CALYOU. Après avoir supprimé les entrées en double, le lexique final contient 125k mots et 538k variantes de prononciation.

4.7 Résultats et discussion

Pour montrer la différence entre l'arabe standard et le dialecte algérien au niveau linguistique et acoustique, nous avons mesuré les performances de reconnaissance de la parole sur la partie test du corpus ADIC (dialecte algérien) avec le système ALASR (appris sur de l'arabe standard). Les résultats en termes du taux d'erreur de mots sont présentés dans le tableau 4.11.

Système	**WER$_{DNN}$**	**WER$_{TDNN}$**	**OOV**
ALASR	78,5%	76,4%	33,6%

TABLE 4.11 – Taux d'erreur mot (WER) et la proportion de mots hors vocabulaire (OOV) mesurés sur la partie test du corpus ADIC avec le système de reconnaissance ALASR développé pour l'arabe standard.

Bien que le système ALASR fonctionne bien sur l'arabe standard, il s'effondre complètement lorsqu'il est appliqué sur le corpus dialectal. Le taux élevé de mots hors vocabulaire, *Out Of Vocabulary (OOV)*, montre à quel point l'arabe standard et le dialecte algérien sont différents. Cela confirme qu'il est impossible de reconnaître directement le dialecte algérien avec un système de reconnaissance automatique de la parole développé pour l'arabe standard.

Nous avons ensuite évalué les différentes approches proposées. Dans un premier temps, un système de base a été développé où tous les modèles ont été entraînés sur les données dialectales seulement. Par la suite, nous avons intégré les données d'autres langues influençant le dialecte (l'arabe standard et le français) dans le processus d'apprentissage. Les différentes configurations sont présentées dans le tableau 4.12 ; pour chaque configuration on indique les données utilisées pour apprendre chaque modèle.

Configurations	**conf1**	**conf2**	**conf3**
Modèle de langage	Dialecte	Dialecte+Ar	Dialecte+Ar
Lexique	Dialecte	Dialecte+Ar	Dialecte+Ar
Modèle acoustique	Dialecte	Dialecte	Dialecte+autres langues

TABLE 4.12 – Configurations étudiées pour la mise en oeuvre du système de reconnaissance automatique de la parole pour le dialecte algérien.

L'objectif de la configuration *conf2* est d'étudier l'impact des données de l'arabe standard sur la modélisation linguistique du dialecte. Pour la configuration *conf3*, on étudie l'apport des données d'autres langues sur la modélisation acoustique du dialecte.

Les résultats de reconnaissance du dialecte selon les deux premières configurations sont rapportés dans le tableau 4.13. Pour la *conf3*, nous détaillons les résultats obtenus en fonction des techniques apprentissage proposées (voir la section 4.4).

Configuration	**Données d'apprentissage**		**WER**	**OOV**
	Linguistiques	**Acoustiques**		
Système ALASR	Ar	Ar	76,4%	33,6%
Conf1	Dialecte	Dialecte	42,6%	7,9%
Conf2	Dialecte+Ar	Dialecte	39,7%	6,8%

TABLE 4.13 – Taux d'erreur mot obtenus sur la partie test du corpus ADIC avant et après l'intégration des données de l'arabe standard dans le lexique et dans le processus d'apprentissage du modèle de langage (données d'apprentissage linguistiques).

On peut remarquer à partir des résultats du tableau 4.13 que l'utilisation des données de l'arabe standard réduit le taux d'erreur mot de 2,9% (en absolu). Cela s'explique, d'une part, par la diminution du taux de mots hors vocabulaire suite à l'intégration de mots de l'arabe standard dans le lexique, et d'autre part, par l'augmentation de la quantité de données textuelles utilisées pour l'apprentissage du modèle de langage.

4.7.1 Apprentissage multilingue

Cette approche vise à tirer profit des données acoustiques d'autres langues pour améliorer la reconnaissance du dialecte algérien en apprenant le réseau de neurones sur un mélange de données du dialecte, de l'arabe standard et du français.

Pour estimer la quantité optimale de données acoustiques de chaque langue à utiliser lors de l'apprentissage, en complément des données du dialecte algérien, nous présentons dans la figure 4.2 l'évolution du taux d'erreur mot sur la partie validation du corpus ADIC en ajoutant à chaque étape 4 heures de données de l'arabe standard et du française. Le chiffre au-dessus de chaque courbe représente la quantité de données de l'arabe standard (en heures). Les courbes en bleu

et en noir représentent respectivement l'évolution du taux d'erreur mot lors de l'utilisation de l'union de phonèmes et lors de l'utilisation de phonèmes communs (voir la section 4.4).

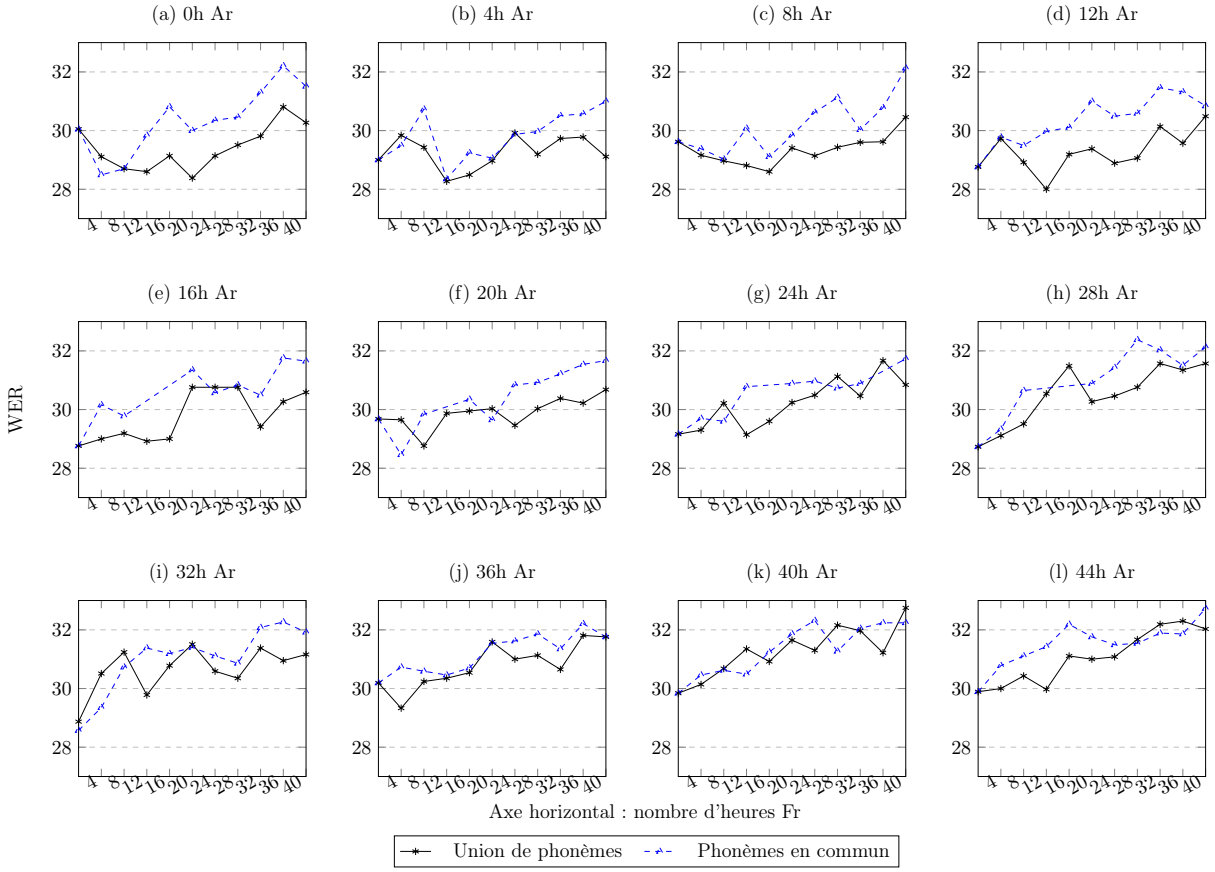


FIGURE 4.2 – Évolution des taux d'erreur mot (WER) sur la partie de validation du corpus ADIC en intégrant plus ou moins de données de l'arabe standard et du français dans le processus d'apprentissage du modèle acoustique.

Le taux d'erreur mot dans le système de base (sans utiliser de données de l'arabe standard ni celles du français) est de 30,05%. Le meilleur résultat (un taux d'erreur mot de 28%) est celui obtenu en ajoutant 12 heures d'arabe standard et 12 heures de français (voir la courbe (d)).

Les expériences montrent que lorsqu'on augmente considérablement la quantité de données d'arabe standard et de français, les résultats de reconnaissance du dialecte se dégradent. Ce comportement était attendu, car l'utilisation de données déséquilibrées entre langues peut rendre le modèle acoustique moins pertinent pour la (ou les) langues les moins bien représentées dans le corpus d'apprentissage. En outre, ces expériences nous ont permis de connaître la quantité exacte de données nécessaires à l'amélioration du modèle acoustique pour le dialecte algérien (ou plus précisément, pour le corpus d'apprentissage du dialecte algérien dont nous disposons).

En ce qui concerne le choix des unités acoustiques pour l'apprentissage du modèle multilingue, la méthode reposant sur l'union des phonèmes est plus performante que celle exploitant le partage des phonèmes communs entre les langues dans la plupart des expériences. Cela peut s'expliquer par le fait que les phonèmes en commun entre l'arabe standard et le français, même s'ils sont identiques, sont utilisés dans des contextes phonologiques différents, ce qui rend leurs prononciations différentes dans les deux langues. Par conséquent, leur modélisation séparée permet de

mieux tenir compte des contextes spécifiques à chaque langue.

Les résultats de reconnaissance sur la partie test du corpus ADIC sont présentés dans le tableau 4.14.

Configuration	**Données d'apprentissage**		**WER**	**OOV**
	Linguistiques	**Acoustiques**		
Conf2	Dialecte+Ar	Dialecte	39,7%	6,8%
Conf3		Dialecte+44hAr	36,6%	
		Union Dial+44hAr+40hFr	36,3%	
		Partage Dial+44hAr+40hFr	37,1%	
		Union Dial+12hAr+12hFr	**35,9%**	
		Partage ADIC+12hAr+12hFr	38,5%	

TABLE 4.14 – Résultats obtenus sur la partie test du corpus ADIC avant et après l'intégration des données de l'arabe standard et du français dans le processus d'apprentissage multilingue du modèle acoustique.

Le système de la *conf2* où seulement les 4 heures de données dialectales sont utilisées pour apprendre le modèle acoustique atteint un taux d'erreur de 39,7%. En étendant ce corpus avec 44 heures d'arabe standard, on obtient une amélioration absolue de 3,1%. Mieux encore, l'intégration des données du français dans le processus d'apprentissage avec l'optimisation de la quantité de données de chaque langue améliore le système de la *conf2* de 3,8%. Il est à noter que l'intervalle de confiance du système de la *conf2* est de ±1,65% ce qui signifie que notre meilleur système (celui qui atteint un taux d'erreur de 35,9%) conduit à une amélioration significative par rapport au système de base (celui de la *conf2*).

Notre conclusion des résultats de cette première approche confirme qu'en alimentant le modèle avec une quantité optimisée de données de plusieurs langues assure de meilleurs résultats de reconnaissance pour le dialecte algérien. En outre, le choix des unités acoustiques joue un rôle primordial dans la modélisation des phonèmes en commun entre les langues ; la modélisation séparée des unités acoustiques de chaque langue assure de meilleurs résultats.

4.7.2 Apprentissage multitâche

Dans le but de capter les similarités entre les langues et de mieux reconnaître le dialecte algérien, nous visons dans cette approche à entraîner un seul modèle acoustique pour reconnaître plusieurs langues en même temps. Dans un premier temps, le modèle est entraîné sur deux tâches, à savoir la reconnaissance du dialecte et de l'arabe standard. Dans un deuxième temps, nous intégrons une nouvelle tâche dans le processus d'apprentissage, celle de la reconnaissance du français. L'objectif de cette démarche est d'étudier l'apport de chaque langue (l'arabe standard et le français) sur la modélisation acoustique du dialecte algérien.

La première approche proposée pour estimer les paramètres du réseau de neurones consiste à attribuer des poids pour chaque langue afin de donner plus d'importance au dialecte algérien (voir section 4.4.2). La réussite de cette approche dépend du bon choix des poids p_l pour chaque langue. Dans le tableau 4.15, nous donnons les taux d'erreur mot obtenus sur la partie de validation du corpus ADIC en variant les poids p_l. Comme notre tâche principale est la reconnaissance du dialecte algérien, alors notre idée pour l'estimation des p_l est de commencer l'apprentissage avec un poids élevé pour le dialecte et de le diminuer progressivement avec un pas de 0,1 de sorte que le dialecte ait toujours le plus grand poids. Nous avons opté pour cette solution car la tâche

de l'apprentissage du réseau de neurones est lente et il est difficile d'explorer tout l'espace de recherche. Il apparaît que l'utilisation des poids quasi-équiprobable entre les différentes langues assure les meilleurs résultats.

(a) Trois tâches de reconnaissance

p_{Dial}	p_{Ar}	p_{Fr}	WER (%)
0,8	0,15	0,05	30,5
	0,1	0,1	29,8
0,7	0,15	0,15	31,4
	0,2	0,1	31,6
0,6	0,2	0,2	31,7
	0,3	0,1	31,6
0,5	0,25	0,25	31,3
	0,4	0,1	31,8
0,4	0,3	0,3	**28,6**

(b) Deux tâches de reconnaissance

p_{Dial}	p_{Ar}	WER (%)
0,9	0,1	30,5
0,8	0,2	30,6
0,7	0,3	30,3
0,6	0,4	29,9
0,5	0,5	**29**

TABLE 4.15 – Variations du taux d'erreur mot sur la partie validation du corpus ADIC selon les poids de pondération p_l de chaque langue.

La deuxième approche pour estimer les paramètres du réseau de neurones consiste à utiliser des mini-batchs incluant des données de chacune des langues. Les résultats de reconnaissance sur la partie test du corpus ADIC sont présentés dans le tableau 4.16.

Configuration	**Données d'apprentissage**		**WER**	**OOV**
	Linguistiques	**Acoustiques**		
Conf2	Dialecte+Ar	Dialecte	39,7%	6,8%
Conf3		Dial+Ar+Fr (apprentissage multilingue)	35,9%	
		Dial+Ar (apprentissage multitâche)	37,0% (poids pondération)	
			36,5% (mini-batch)	
		Dial+Ar+Fr (apprentissage multitâche)	36,9% (poids pondération)	
			36,6% (mini-batch)	

TABLE 4.16 – Taux d'erreur mot obtenu sur la partie test du corpus ADIC avec l'apprentissage multitâche.

Les résultats du tableau 4.16 montrent que la meilleure technique pour estimer les paramètres du modèle acoustique est celle où aucun poids n'est associé aux différentes tâches de reconnaissance. Les données de différentes langues ont, dans ce cas, la même importance dans le processus d'estimation de paramètres du réseau de neurones. Cela pourrait être expliqué par notre approche d'estimation des poids de pondération de chaque langue qui explore un espace de recherche très restreint. L'autre remarque importante concerne l'impact de données de la langue française sur la reconnaissance du dialecte. L'intégration de la tâche de reconnaissance du français dans le processus d'apprentissage multitâche n'améliore pas la reconnaissance du dialecte algérien. Cela est principalement dû à la relation qui existe entre les différentes tâches. Il est connu que l'apprentissage multitâche fonctionne bien dans le cas où les tâches sur lesquelles on apprend notre modèle sont proches [Zhang et Yang, 2017]. La décision de dire que les tâches sont liées et qu'elles peuvent être apprises par un seul modèle est généralement prise par des experts humains et il existe peu d'approches automatique pour l'étudier. Dans notre cas, l'intégration de

la tâche de reconnaissance automatique de la parole française n'améliore pas la reconnaissance du dialecte.

4.7.3 Transfert de connaissances

En partant de l'hypothèse que les couches cachées d'un réseau neuronal sont génériques et que la dernière est spécifique à une tâche donnée, nous avons entraîné un modèle initial sur les données de l'arabe standard et du français en utilisant notre approche d'apprentissage multilingue. Ensuite, les n premières couches cachées (avec $n \in \{1, 2, 3, 4, 5\}$ sachant que le modèle initial est composé de 6 couches cachées) de ce modèle ont été extraites et au-dessus, une nouvelle couche de sortie spécifique à la tâche de reconnaissance du dialecte est ajoutée. Rappelons que les paramètres des couches cachées sont réajustés avec un *learning rate* faible et les paramètres de la dernière couche spécifique au dialecte sont estimés avec un *learning rate* plus grand. Les résultats obtenus en termes du taux d'erreur mot sur la partie validation du corpus ADIC sont présentés dans la figure 4.3. Nous remarquons que le fait de garder un maximum de couches cachées du modèle initial et de les transférer pour le nouveau modèle du dialecte assure de meilleurs résultats. Les tests réalisés sur la partie du test du corpus ADIC et présentés dans le tableau 4.17 sont effectués en gardant les cinq premières couches cachées du modèle initial, en réajustant les paramètres de ces couches, et en ajoutant une nouvelle couche spécifique au dialecte.

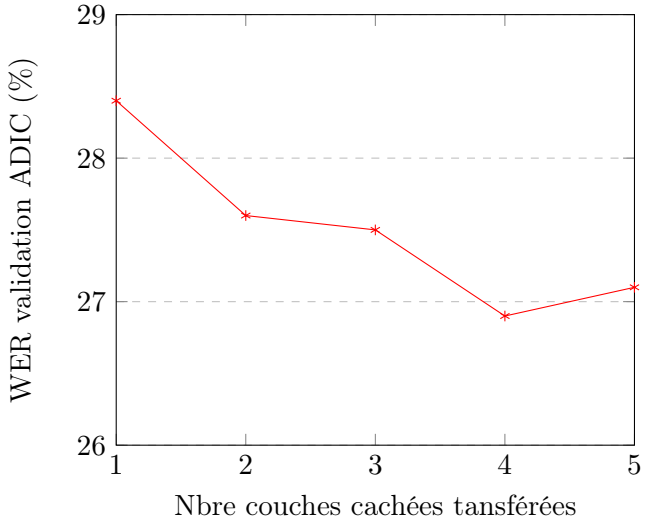


FIGURE 4.3 – L'impact du nombre de couches cachées transférées au modèle dialectal.

Le tableau 4.17 récapitule nos résultats obtenus pour la reconnaissance automatique de la parole pour le dialecte algérien.

Technique d'apprentissage	Données d'apprentissage		WER
	Linguistiques	**Acoustiques**	
Apprentissage monolingue	Ar	44hAr	76,3%
	Dial	4hDial	42,6%
	Dial+Ar	4hDial	39,7%
Apprentissage multilingue	Dial+Ar	44hAr+40hFr	47,6%
		4hDial+44hAr	36,6%
		4hDial+44hAr+40hFr	36,3%
		4hDial+12hAr+12hFr	**35,9%**
Apprentissage multitâche		4hDial+44hAr	36,5%
		4hDial+44hAr+40hFr	36,6%
Transfert de connaissances		44hAr modèle initial=>4hDial	38,1%
		44hAr+40Fr modèle initial=>4hDial	37,1%

TABLE 4.17 – Résultats comparatifs de la reconnaissance automatique de la parole sur le corpus de test en fonction de l'approche d'intégration des données de langues étrangères dans le processus d'apprentissage des modèles acoustique et de langage.

En comparant nos trois approches d'intégration de données de plusieurs langues dans le processus d'apprentissage du modèle acoustique du dialecte algérien, nous avons trouvé que la meilleure approche est celle basée sur l'apprentissage multilingue. Dans ce cas, toutes les couches du réseau de neurones sont partagées entre les trois langues. Cela permet une augmentation implicite de la taille des données que nous utilisons pour entraîner notre modèle acoustique, ce qui permet à ce dernier de mieux capturer la relation entre les trois langues et d'améliorer, par conséquent, la reconnaissance du dialecte.

Il est à noter que les travaux de recherche sur la reconnaissance automatique de la parole pour le dialecte algérien sont relativement peu nombreux. Cependant, dans la dernière édition de la compétition MGB, MGB5 [Ali *et al.*, 2019], il y avait une tâche sur la reconnaissance automatique de la parole pour le dialecte marocain. Ce dernier est relativement proche du dialecte algérien, ils partagent plusieurs aspects linguistiques et acoustiques. Le meilleur système a obtenu un taux d'erreur de 37,6%, sachant que 13 heures de la parole dialectale ont été utilisées avec 1200 heures de l'arabe standard pour apprendre le modèle acoustique. Cela montre la difficulté de reconnaître les dialectes maghrébins et que les résultats de notre système sont acceptables.

4.8 Conclusion et discussion

Nous avons présenté dans ce chapitre notre travail sur l'adaptation du système ALASR pour le dialecte algérien. Ce dernier est l'une des variantes de la langue arabe les plus difficiles à reconnaître par les systèmes de reconnaissance automatique de la parole. Cela est principalement dû aux mots empruntés d'autres langues qui tiennent une place importante dans le vocabulaire du dialecte algérien. Nous avons illustré à quel point le dialecte algérien et l'arabe standard sont différents en montrant l'impossibilité de reconnaître le dialecte algérien avec notre système ALASR développé pour l'arabe standard.

Le problème du dialecte algérien est le manque de ressources vocales pour apprendre le modèle acoustique. Notre approche pour remédier à ce problème était de tirer profit des données orales d'autres langues qui ont un impact sur le dialecte, à savoir l'arabe standard et le français. Nous avons montré qu'en enregistrant un petit corpus dialectal (de seulement 4 heures), il est possible de développer un modèle acoustique en utilisant en complément des données de langues bien

dotées en ressources telles que le français et l'arabe standard. Il pourrait être intéressant d'étudier cette approche pour développer des systèmes de reconnaissance automatique de la parole pour d'autres dialectes, en particulier pour ceux qui sont influencés par le français comme les dialectes marocains et tunisiens. Le corpus dialectal enregistré constitue une ressource précieuse pour des études plus approfondies sur le dialecte algérien. Il est disponible en libre accès.

Nous avons trouvé que le meilleur moyen pour exploiter les données d'autres langues dans la modélisation acoustique du dialecte est d'apprendre le réseau de neurones du modèle acoustique sur un mélange de données de plusieurs langues tout en partageant toutes les couches du réseau de neurones entre les différentes langues. Dans ce cas, le choix des unités acoustiques à modéliser est primordial vu que nos trois langues partagent des phonèmes entre eux. Modéliser les phonèmes de chaque langue séparément est la meilleure solution dans le cas de la reconnaissance du dialecte algérien. En outre, nous avons montré qu'en sélectionnant un sous-ensemble de données de chaque langue, cela conduit à un meilleur modèle acoustique par rapport à l'utilisation de toutes les données disponibles (tant en arabe qu'en français). Cette investigation nous a permis d'éviter le cas de sur-apprentissage du modèle sur une langue autre que le dialecte algérien.

Chapitre 5

Approche statistique vs. neuronale dans la traduction automatique

Sommaire

Dans ce chapitre, nous présentons nos contributions dans le cadre de la traduction automatique. Notre première contribution consiste à faire une étude comparative entre l'approche statistique, la plus utilisée ces 20 dernières années, et l'approche neuronale qui fait ses preuves dans le domaine de la traduction automatique. Cette étude est réalisée dans le cas de la traduction arabe-anglais où on suppose que peu de données d'apprentissage sont disponibles sachant que l'approche neuronale est connue dans la littérature pour être gourmande en ressources. Cette supposition est faite pour étudier le comportement de chaque approche dans un cadre où peu de données sont disponibles afin d'avoir une estimation de ce même comportement vis-à-vis du dialecte. Nous nous concentrons sur quelques techniques en particulier pour booster le système de traduction basé sur l'approche neuronale afin d'avoir des résultats comparables à l'approche statistique [Menacer *et al.*, 2017a].

Nous commençons par présenter le contexte général de notre étude comparative. Nous présentons ensuite les données utilisées. Enfin, nous discutons les aspects qui peuvent améliorer la traduction neuronale notamment dans le cas où peu de données d'apprentissage sont disponibles.

5.1 Contexte de l'étude

Dans le chapitre 2, nous avons présenté plusieurs approches qui ont été proposées dans la littérature pour automatiser le processus de la traduction. Certaines d'entre elles sont basées sur les dictionnaires [Hutchins, 2004b], d'autres sur des exemples [Somers, 1999], des règles linguistiques [Dugast *et al.*, 2007] ou sur des approches statistiques [Brown *et al.*, 1993]. La technique la plus utilisée ces 20 dernières années est la traduction automatique statistique basée sur les segments [Zens *et al.*, 2002]. Cette approche est basée sur la même idée que celle appliquée dans les systèmes de reconnaissance automatique de la parole, notamment l'utilisation de deux modèles celui de la traduction et de langage pour maximiser la probabilité de traduire une phrase source F en une phrase cible E. Le modèle de traduction, entraîné sur un corpus parallèle, exprime à quel point la phrase E est une traduction appropriée pour la phrase source F. Le modèle de langage est entraîné sur un corpus monolingue afin d'assurer que la phrase cible E soit bien écrite dans la langue cible.

Ces dernière années, l'apprentissage basé sur les réseaux de neurones est devenu une technique de référence largement utilisée dans plusieurs domaines de recherche, notamment la reconnaissance automatique de la parole, la reconnaissance d'objets visuels, l'analyse des sentiments, etc. La théorie derrière ces approches n'est pas récente, elle est établie depuis plusieurs décennies, mais son utilisation s'est répandue ces dernières années grâce à la disponibilité des données d'apprentissage et à la puissance des machines. Compte tenu de ces facteurs, il n'est pas surprenant que les approches neuronales aient récemment déclenché une tempête dans le domaine de la traduction automatique pour créer une approche prometteuse appelée la traduction automatique neurale, *Neural Machine Translation (NMT)* [Kalchbrenner et Blunsom, 2013, Sutskever *et al.*, 2014, Cho *et al.*, 2014a].

Contrairement à l'ancien paradigme des approches statistiques de la traduction automatique où on devait explicitement modéliser les structures latentes de la langue, à savoir : l'alignement des mots, la segmentation des phrases, le réordonnancement des phrases et la modélisation du langage, le nouveau paradigme basé sur les modèles neuronaux est un modèle de bout en bout. Il vise à transformer directement une phrase source en une phrase cible avec un seul réseau de neurones. Ce dernier est basé sur deux composants : l'encodeur et le décodeur (voir section 2.3 pour plus de détails sur le fonctionnement de ces deux composants).

Plusieurs analyses comparatives entre l'approche statistique et l'approche neuronale ont été élaborées [Isabelle *et al.*, 2017, Koehn et Knowles, 2017, Bentivogli *et al.*, 2016], afin d'explorer les défis du nouveau paradigme et de bien comprendre quels phénomènes linguistiques sont les mieux modélisés par les modèles neuronaux. Ces analyses ont été principalement menées sur des paires de langues latines. Les deux principales différences entre ces analyses et notre étude présentée dans ce chapitre sont les suivants :

— La paire de langues étudiée : nous nous concentrons dans notre travail sur la traduction de l'arabe vers l'anglais. La motivation derrière ce choix est le peu de travaux de recherche sur le comportement de chaque approche de traduction sur la paire de langues arabe-anglais. Ainsi, nos travaux de thèse font partie du projet AMIS dont l'un de ces objectifs est la traduction de vidéos de l'arabe vers l'anglais.

— La taille des données d'apprentissage : dans nos travaux de recherche, on se focalise sur le traitement automatique de langues peu dotées en ressources, en particulier le dialecte algérien. Or, les corpus parallèles dialectaux sont peu disponibles, donc nous avons décidé de mener cette étude comparative sur un corpus arabe-anglais de taille réduite. Nous étudions en particulier comment on peut booster le modèle neuronal pour atteindre des résultats état de l'art, sachant que ces modèles sont gourmands en ressources.

5.2 Données utilisées

Nos données font partie du corpus parallèle arabe-anglais de MultiUN [Eisele et Chen, 2010]. Le corpus MutliUN est une collection de documents des nations unies traduits dans les six langues officielles des nations unies, à savoir l'anglais, l'arabe, le chinois, l'espagnol, le français et le russe. Ces documents correspondent à la période comprise entre janvier 2000 et septembre 2009. Le corpus parallèle arabe-anglais de MultiUN est un large corpus contenant plus de 9,7M de phrases.

Dans le cadre de notre étude comparative, nous avons sélectionné à partir du corpus global un petit corpus de 100k phrases parallèles pour l'apprentissage, 1k phrases pour la validation et de même pour le test. La sélection est faite après une phase de normalisation de données. La normalisation repose sur le processus proposé et mis en oeuvre pour normaliser les données textuelles du système ALASR. En outre, nous avons ajouté une étape complémentaire de traitement qui consiste à :

— Garder seulement les phrases dont le nombre de mots varie entre 2 et 50 mots.
— *Tokeniser* les données afin de les découper en des unités plus petites (des mots dans notre cas).
— *Truecasing* qui consiste à déterminer la bonne capitalisation des mots. Cette dernière étape est appliquée uniquement pour le texte en anglais car la notion de majuscule et de minuscule n'existe pas en arabe.

Le tableau 5.1 donne quelques informations sur la taille des corpus parallèles utilisés pour comparer les deux approches statistique et neuronale sur la paire de langue arabe-anglais.

Partie	**Langues**	**Nombre de phrases**	**Nombre de mots**	**Nombre de mots uniques**
Apprentissage	Ar	100k	2,2M	81k
	En		2,6M	31k
Validation	Ar	1k	21k	5,5k
	En		24k	3,2k
Test	Ar	1k	21k	5,9k
	En		26k	3,1k

TABLE 5.1 – Informations sur les corpus d'apprentissage, de validation et de test.

Quelque soit la taille des corpus d'apprentissage, il existe toujours des mots inconnus de l'ensemble du test qui n'apparaissent pas dans le corpus d'apprentissage. Le moyen le plus simple pour traiter ces mots est de remplacer les mots les moins fréquents (ceux qui n'apparaissent qu'une seule fois dans le corpus d'apprentissage dans notre cas) par un mot spécial <*unk*>. Ce processus nous a permis d'obtenir un vocabulaire de 47K mots pour la langue source arabe et de 20K mots pour la langue cible anglais. Nous pouvons remarquer que la taille du vocabulaire de l'anglais ainsi que le nombre de mots uniques sont plus faibles par rapport à l'arabe. Cela est principalement dû à la nature agglutinante de la langue arabe.

En se basant sur ce corpus, nous avons comparé les approches statistique et neuronale selon les aspects suivants : la modélisation de l'alignement des mots de la phrase source et ceux de la phrase cible, la traduction des mots hors vocabulaire et le processus de décodage.

5.3 Systèmes de base

Afin de pouvoir comparer les approches statistique et neuronale selon les aspects cités ci-dessus, nous avons développé deux systèmes de base, un pour chaque approche. Notre étude com-

mence par la comparaison de ces deux systèmes de base en termes de BLEU [Papineni *et al.*, 2002] pour passer, par la suite, aux approches qui visent à booster le système neuronal.

5.3.1 Approche statistique à base de segments

Deux composants sont nécessaires pour le développement du système de traduction à base de segments, à savoir : le modèle de traduction et le modèle de langage (voir section 2.2). Ce dernier est un modèle 3-grammes que nous avons entraîné sur le corpus de la langue anglaise avec l'outil SRILM [Stolcke *et al.*, 2011]. Le modèle de traduction est entraîné avec l'outil GIZA++ [Och et Ney, 2003] et le décodage est basé sur l'outil MOSES [Koehn *et al.*, 2007a]. Lors du processus du décodage, le score final d'une traduction e est calculé en fonction de sept scores h_i comme le montre l'équation 5.1.

$$score(e) = \prod_{i=1}^{7} h_i(f, e)^{\lambda_i} \tag{5.1}$$

Les score h_i sont :
— h_1 : score du modèle de langage $P(e)$.
— h_2 : score du modèle de traduction directe $P(e|f)$.
— h_3 : score du modèle de traduction inverse $P(f|e)$.
— h_4 : score du modèle lexical direct $P(e_i|f_j)$ [13] (avec $1 \leq i \leq |e|$ et $1 \leq j \leq |f|$).
— h_5 : score du modèle lexical inverse $P(f_j|e_i)$.
— h_6 : score de réordonnancement.
— h_7 : pénalité de mots pour contrôler la taille de la traduction.

Les poids λ_i sont associés aux score h_i pour préciser l'importance de chaque modèle dans le calcul du score final. L'optimisation de ces poids est réalisée sur la partie de validation de notre corpus avec l'algorithme de *Minimum Error Rate Training (MERT)* [Och, 2003].

5.3.2 Approche neuronale

Le système neuronal est un modèle à base d'encodeur-décodeur sans aucun mécanisme d'alignement comme il a été proposé dans [Cho *et al.*, 2014b, Sutskever *et al.*, 2014] (voir section 2.3). Pour ce système, l'architecture utilisée est la suivante : une couche cachée dans laquelle nous avons utilisé des cellules RNN pour coder les phrases sources ainsi que pour les décoder. Chaque cellule comporte 100 unités cachées, ce qui signifie que la dimension des vecteurs générés par chaque cellule RNN est de 100. En outre, plusieurs algorithmes d'optimisation ont été testés afin de mettre à jour les paramètres du modèle, à savoir :

Stochastic Gradient Descent (SGD) Cette technique est basée sur le calcul de la dérivée de la fonction de perte par rapport à chaque paramètre à estimer. Cette dérivée est utilisée par la suite, pour mettre à jour les paramètres en se déplaçant dans la direction opposée au gradient. Le pas de déplacement est défini par l'hyperparamètre *learning rate*.

Momentum [Qian, 1999] Cette technique est basée sur l'algorithme SGD. Elle est utilisée pour aider l'*optimiseur* à explorer plus efficacement l'espace des paramètres en gardant une fraction du gradient passé et en l'ajoutant au gradient actuel. Avec cette technique, on assure une convergence plus rapide du processus d'optimisation et on réduit l'oscillation lors de l'exploration de l'espace de recherche des paramètres.

13. Cette probabilité estime la qualité de l'alignement entre les mots du segment source et ceux du segment cible.

AdaGrad [Duchi *et al.*, 2011] Cette technique est très utile dans le cas de données clair-semées. En effet, elle est basée sur l'ajustement dynamique du *learning rate* pour chaque paramètre de sorte que des pas de mise à jour plus importants sont effectuées pour les paramètres peu fréquemment modifiés et des pas de mise à jour plus petits pour les paramètres fréquemment modifiés.

Adaptive moment estimation (Adam) [Kingma et Ba, 2014] Il s'agit d'une autre technique qui calcule des *learning rates* adaptatifs pour les différents paramètres. Elle combine les avantages du *Momentum* et d'*AdaGrad* en conservant une fraction des gradients passés (moyenne et variance) et en les ajoutant aux gradients actuels. C'est une technique d'optimisation populaire car sa convergence est très rapide mais nous avons décider de la comparer avec la méthode *SGD*.

5.4 Alignement

Dans le modèle neuronal de base, la séquence de mots cible est générée à partir de la représentation intermédiaire de la phrase source. Il n'y a aucun mécanisme pour modéliser l'alignement entre les mots de la phrase source et ceux de la cible avec ce modèle. La solution pour ce problème est les techniques d'*attention* [Bahdanau *et al.*, 2014]. Dans ce cas, la phrase source est codée avec un vecteur numérique appelé un vecteur de contexte qui est calculé dynamiquement en fonction des mots cibles à générer (voir section 1.2.2 pour plus de détails). Avec cette approche, le modèle arrive à définir pour chaque mot de la phrase cible E les mots les plus importants de la phrase source F qui peuvent générer ce mot cible. À la fin, on aura une estimation des probabilités d'aligner chaque mot de la phrase cible avec ceux de la phrase source.

Une autre approche intéressante qui permet de mieux coder la phrase source et de mieux générer par conséquent, la phrase cible est l'encodage bidirectionnel. Dans l'architecture de base, l'encodage de la phrase source F commence par le premier mot f_1 jusqu'au le dernier $f_{|F|}$. Par contre dans l'encodage bidirectionnel, la phrase d'entrée est, tout d'abord, codée du premier mot jusqu'au dernier pour générer une séquence d'états cachés $(\overrightarrow{h_1^{(f)}}, ..., \overrightarrow{h_{|F|}^{(f)}})$. Elle est ensuite codée dans l'ordre inverse, ce qui nous permet d'obtenir une autre séquence d'états cachés $(\overleftarrow{h_1^{(f)}}, ..., \overleftarrow{h_{|F|}^{(f)}})$. Enfin, pour obtenir le vecteur h_i qui correspond à chaque mot f_i, les deux états cachés de sortie $\overrightarrow{h_i^{(f)}}$ et $\overleftarrow{h_i^{(f)}}$ sont concaténés comme suit $h_i^{(f)} = [\overrightarrow{h_i^{(f)}}; \overleftarrow{h_i^{(f)}}]^T$ [14].
Avec cette approche, les valeurs des états cachés associés à chaque mot f_i dépendent à la fois des valeurs des mots précédents et celles des mots suivants. Cela est très utile dans le cas de paires de langues qui partagent la même structure *sujet-verbe-objet (SVO)* comme la paire de langues français-anglais. En effet, pour ces paires de langues, le mot f_1 est généralement aligné avec le mot e_1, le mot f_2 est aligné avec le mot e_2 et ainsi de suite. Cela signifie qu'en utilisant l'encodage linéaire, la distance entre le mot f_1 et le mot e_1 est égale à $|F|$ et que l'encodeur doit propager l'information sur $|F|$ pas de temps avant de faire une prédiction du mot e_1. Avec cette technique d'encodage et à cause du problème du gradient en voie de disparition, *Vanishing gradient* [Neubig, 2017], on risque de perdre l'information. Avec l'encodage bidirectionnel, cette distance entre les mots de la phrase source et ceux de la phrase cible est réduite.

14. Le transposé du vecteur.

5.5 Traduction des mots hors vocabulaire

L'approche neuronale fonctionne avec un vocabulaire restreint, une grande quantité de mots est remplacée par <*unk*>. L'apprentissage des systèmes neuronaux actuels est généralement basé sur un vocabulaire de 50k mots. Pour certaines langues avec une morphologie plus riche (comme l'arabe), apprendre le modèle avec seulement 50k mots n'est pas efficace. En effet, la phrase cible produite peut contenir plusieurs mots inconnus remplacés par <*unk*> ce qui brise la structure de la traduction et par conséquent en modifie son sens.

Une façon simple pour résoudre ce problème est de tirer profit d'un dictionnaire externe contenant une liste de mots avec leurs traductions ainsi que la probabilité de chaque traduction. Dans notre cas, nous avons examiné deux approches :

RepUnk [Luong *et al.*, 2014] Cette approche consiste à utiliser les probabilités d'alignement (estimées suite au modèle d'attention) pour définir pour chaque mot inconnu dans le corpus de test, les mots de la phrase source qui lui correspondent. Ensuite, ces mots de la phrase source sont traduits avec un dictionnaire externe.

ProbAdjust [Arthur *et al.*, 2016] Cette approche est basée sur l'ajustement des probabilités des mots inconnus lors de la phase d'apprentissage du modèle. Pour ce faire, la probabilité de chaque mot du vocabulaire de la langue cible est recalculée en prenant compte de la probabilité des mots dans le dictionnaire externe. Ainsi, la probabilité d'un mot dans le vocabulaire est ajustée par celle calculée à partir du lexique externe. Pour chaque mot e_i dans le corpus d'apprentissage de la langue cible, on estime sa probabilité conditionnelle lexicale comme suit :

$$p_{lex}(e_i|\{e_1,...,e_{i-1}\},F) = \sum_{j=1}^{|F|} \alpha_{ij} q_{lex}(e_i|f_j) \tag{5.2}$$

avec α_{ij} sont les probabilités d'alignement estimées par le modèle d'attention, q_{lex} sont les probabilités de traduire un mot f_j en e_i selon le dictionnaire externe. Les probabilités p_{lex} est ajoutée comme un biais aux probabilités $softmax$ calculées par le réseau neuronal.

Pour mettre en place ces deux techniques, nous avons construit automatiquement une table de traduction de 10M d'entrées [15] en alignant 9M de phrases parallèles.

5.6 Architecture neuronale

Pour mieux modéliser et apprendre les dépendances à long terme entre les mots de la phrase source ou ceux de la phrase cible, nous avons utilisé des cellules LSTM au niveau de l'encodeur et du décodeur au lieu des cellules RNN. L'avantage principal des cellules LSTM est d'éviter le problème du gradient en voie de disparition, *Vanishing gradient* [Neubig, 2017] dont souffrent les cellules RNN. Nous avons aussi varié le nombre de couches cachées ainsi que la taille des unités cachée. Rappelons que cette taille représente la dimension des vecteurs générés par chaque cellule LSTM.

15. Les entrées de la table de traduction sont une liste de mots avec leurs traductions ainsi que les scores de traduction.

5.7 Décodage

Comme dans l'approche statistique, la production de la phrase cible est basée sur un algorithme de recherche en faisceau. L'idée est d'explorer à chaque étape un sous-ensemble de traductions possibles de taille m (la taille du faisceau). Cette taille a un fort impact sur la qualité de la traduction ; en augmentant la taille du faisceau, le décodeur explore un plus grand sous-ensemble de traductions possibles et par conséquent il assure une meilleure traduction. Lors des expériences précédentes, la taille du faisceau a été fixée à 1, ce que nous considérons comme trop restrictif. En outre, l'évaluation des traductions intermédiaires est effectuée en multipliant les probabilités de générer chaque mot cible sachant les mots déjà générés comme le montre l'équation 5.3.

$$P(E|F) = \prod_{i=1}^{|E|} p(e_i|\{e_1, e_2, ..., e_{i-1}\}, F) \tag{5.3}$$

avec $p(e_i|\{e_1, e_2, ..., e_{i-1}\}, F)$ représentant la probabilité du mot e_i, connaissant les mots précédents, estimée par le réseau de neurones. Dans ce cas, le modèle favorise les séquences de mots les plus courtes car le produit des probabilités de mots devient de plus en plus faible si la taille de la traduction à générer est importante. En contrepartie, en traduisant des phrases de l'arabe vers l'anglais, les traductions générées ont tendance à être plus longues que les phrases sources vu la complexité morphologique de l'arabe ; un seul mot arabe peut être traduit en plusieurs mots en anglais. Une solution à ce problème consiste à intégrer un score de pénalité de mots p_w dans le calcul du score final pour contrôler la longueur des traductions. À chaque fois qu'un mot supplémentaire est traduit, on multiplie la probabilité de la phrase par la constante p_w. Une valeur de $p_w = 1$ signifie qu'aucune modification n'est faite sur la probabilité de la phrase générée. En augmentant la valeur de p_w, on privilégie les traductions longues.

5.8 Résultats et discussion

Dans la figure 5.1, nous présentons les résultats comparatifs entre les techniques d'optimisation utilisées dans le système neuronal de base.

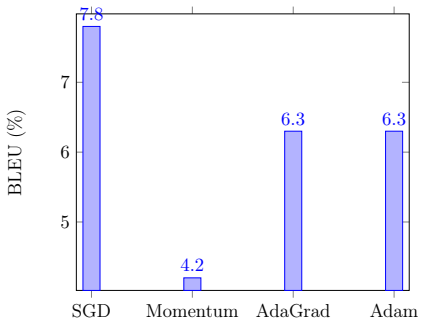


FIGURE 5.1 – Résultats de traduction de la partie de validation en termes de BLEU en fonction de la technique d'optimisation des paramètres du modèle neuronal.

Sur nos données d'apprentissage, l'algorithme d'optimisation SGD nous a permis d'obtenir de meilleurs résultats. Dans la littérature, il est démontré que l'algorithme SGD tend à trouver les paramètres optimaux, mais il souffre de deux problèmes : le temps de convergence qui est plus

élevé par rapport aux autres techniques et le risque de se retrouver coincé dans les points-selle [16] [Ruder, 2016]. Nous pouvons également remarquer qu'*AdaGrad* et *Adam*, qui sont basés sur la même idée, donnent des résultats similaires.

Les résultats de traduction sur la partie de test avec les deux systèmes de base (statistique et neuronal) sont présentés dans le tableau 5.2.

Modèles	**BLEU**	**OOV**
Statistique	24,5%	5,7%
Neuronal	5,4%	

TABLE 5.2 – Performances de traduction comparatives entre l'approche statistique à base de segments (modèle de langage 3-grammes) et l'approche neuronale (une couche cachée, des cellules RNN et 100 unités cachées par cellule).

Dans le cas où peu de données d'apprentissage (100k phrases parallèles) sont utilisées pour entraîner les deux approches, il est clair que le modèle statistique de base assure de meilleurs résultats de traduction en termes de BLEU (24,5% contre 5,35%). En outre, les résultats obtenus avec l'approche neuronale de base sont très mauvais car ils produisent des traductions qui ne sont pas compréhensibles. Dans ce qui suit, nous présentons les résultats des techniques proposées visant à booster le système de traduction neuronal sur nos données d'apprentissage.

5.8.1 Alignement

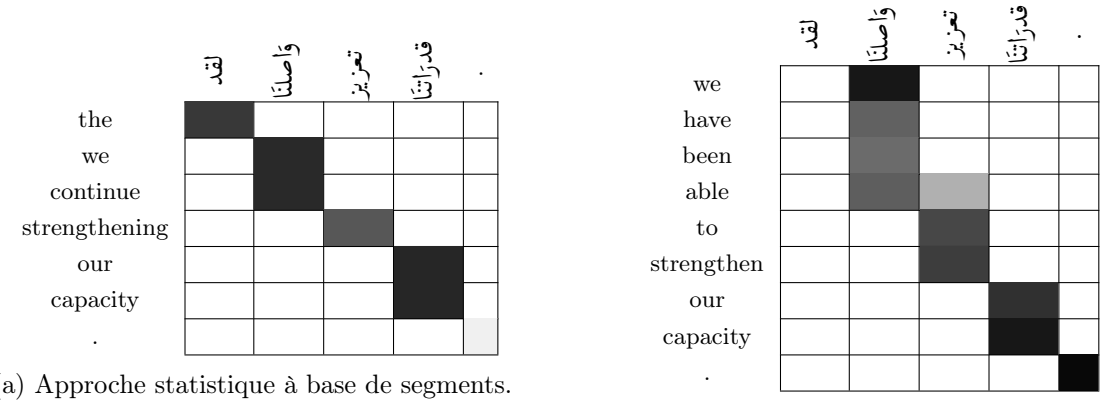
Dans le modèle neuronal, l'alignement entre les mots de la phrase source et ceux de la phrase cible est modélisé via la technique d'attention. En appliquant cette technique et en utilisant l'encodage bidirectionnel pour coder les phrases sources, nous avons obtenu les résultats du tableau 5.3.

Modèles	**BLEU**	**OOV**
Statistique	24,5%	5,7
Neuronal	5,4%	
Neuronal+Attention+Encodage bidirectionnel	20,6%	

TABLE 5.3 – Performances de traduction comparatives entre l'approche statistique à base de segments et l'approche neuronale avec/sans la technique d'attention [Bahdanau *et al.*, 2014] et l'encodage bidirectionnel.

Pour bien comprendre le principe de l'attention, nous donnons dans la figure 5.2 un exemple d'alignement entre la phrase source .لقد وَاصلنَا تعزيز قدرَاتنَا *lqd wāṣlnā tʿzyz qdrātnā.* et ses traductions *(the we continue strengthening our capacity.)* (générée par l'approche statistique) et *(we have been able to strengthen our capacity)* (générée par l'approche neuronale).

16. Des points où la fonction objective atteint une valeur minimale pour certains paramètres et une valeur maximale pour d'autres paramètres.



(a) Approche statistique à base de segments.

(b) Approche neuronale.

FIGURE 5.2 – Exemple comparatif entre l'alignement obtenu par l'approche statistique et l'approche neuronale. L'intensité du gris de chaque case représente les probabilités d'alignement ; plus la case est noire meilleure est la probabilité.

L'intégration de la technique d'attention dans la modélisation basée sur le réseau de neurones a engendré une amélioration notable dans les résultats de la traduction. Nous pouvons remarquer à partir de l'exemple de la figure 5.2 que le modèle neuronal arrive bien à aligner les mots de la phrase source avec ceux de la phrase cible. Une autre remarque importante concerne la structure de la phrase cible générée ; avec l'approche neuronale, on arrive à produire une phrase plus correcte. La traduction générée avec l'approche statistique est correcte à l'exception de l'article *the* au début de la phrase qui n'est pas à sa place. Cela peut être expliqué par le processus de construction de la phrase cible. Dans l'approche statistique, la phrase source est décomposée en plusieurs segments et la phrase cible est construite d'une manière incrémentale en traduisant à chaque étape un segment de la phrase source. Par ailleurs, le modèle neuronal construit la phrase cible en se basant sur la représentation de la phrase source entière, ce qui lui permet de mieux capturer le contexte et la relation entre les mots.

5.8.2 Traduction des mots hors vocabulaire

Nous présentons dans le tableau 5.4 les résultats obtenus en traitant les mots hors vocabulaire en remplaçant les mots inconnus (RepUnk) ou en ajustant les probabilités conditionnelles estimées par le réseau de neurones (ProbAdjust).

Modèles	**BLEU**	**OOV**
Statistique	24,5%	5,7%
Neuronal	5,4%	
Neuronal+Attention+Encodage bidirectionnel	20,6%	
Neuronal+Attention+Encodage bidirectionnel+RepUnk	21,1%	
Neuronal+Attention+Encodage bidirectionnel+ProbAdjust	19,8%	

TABLE 5.4 – Performances de traduction comparatives entre l'approche statistique à base de segments, l'approche neuronale avec/sans la technique d'attention et en traitant les mots hors vocabulaire.

L'utilisation d'un lexique externe pour remplacer les mots inconnus avec leur traduction améliore légèrement la qualité de la traduction sur la partie de test. Cependant, en ajustant

les probabilités de traduction, les performances du système diminuent sur notre ensemble de données. Cela s'explique par le fait que les probabilités externes du lexique peuvent améliorer la probabilité de segments peu fréquents, mais elles peuvent aussi faire chuter les probabilités de ceux qui sont fréquents.

5.8.3 Architecture neuronale

Les résultats de traduction en utilisant les blocs LSTM, en variant la taille des unités cachée et en augmentant le nombre de couches cachées sont présentés dans la figure 5.3.

Sur nos données d'apprentissage, l'utilisation d'une architecture plus profonde n'améliore pas les performances de la traduction neuronale. Nous pouvons expliquer ce constat par le fait qu'une architecture plus profonde est souvent utilisée dans le cas où une grande quantité de données d'apprentissage est disponible afin de mieux extraire les caractéristiques linguistiques et de les apprendre efficacement.

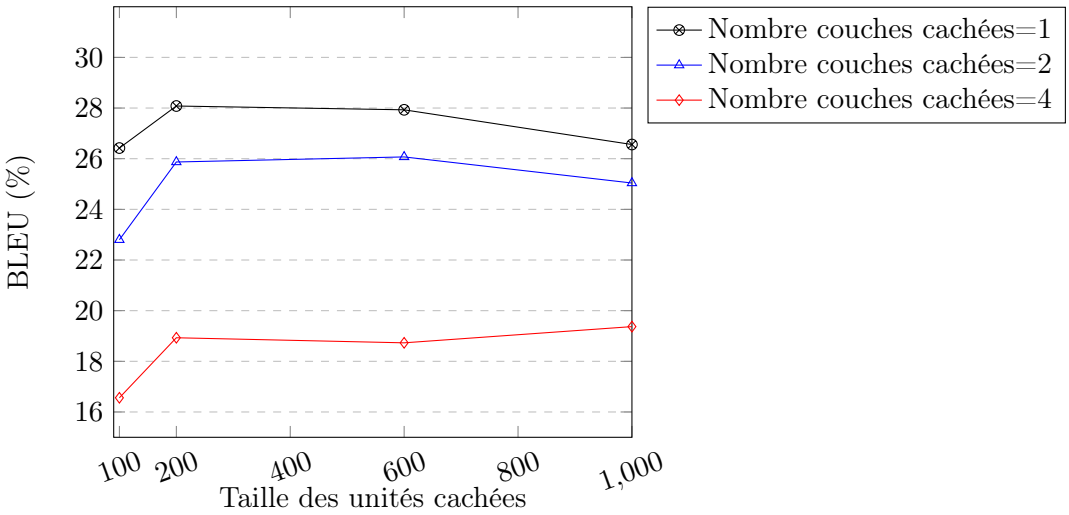


FIGURE 5.3 – Résultats de traduction sur le corpus de validation en utilisant des blocs LSTM pour coder/décoder les phrases source/cible, en variant la taille des unités cachées et en augmentant le nombre de couches cachées.

5.8.4 Décodage

Les courbes de la figure 5.4 présentent les résultats de la traduction du corpus de validation en fonction de la taille du faisceau et de la pénalité de mots p_w.

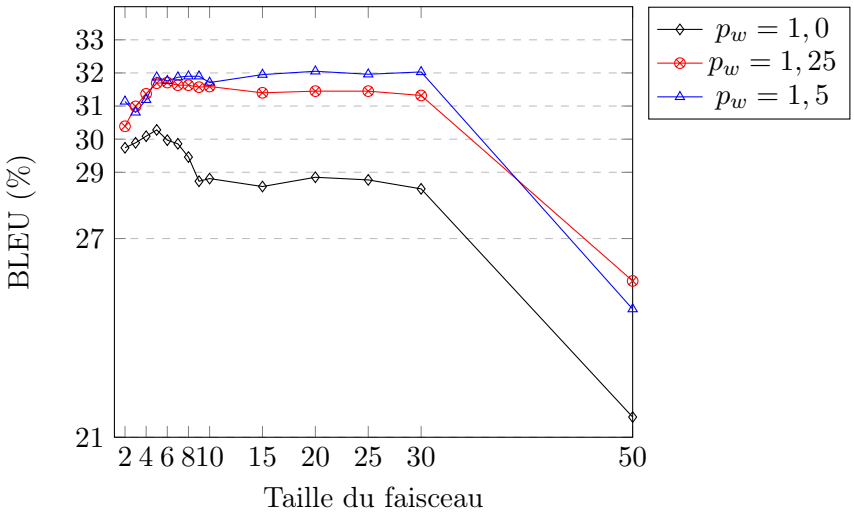


FIGURE 5.4 – Résultats de la traduction du corpus de validation en fonction de la taille du faisceau et de la pénalité de mots p_w.

Les résultats présentés dans la figure 5.4 montrent qu'au-delà d'une valeur optimale de la taille du faisceau (entre 20 et 30), les traductions sont plus mauvaises. Ainsi, en fixant la pénalité de mots à 1,5 (la valeur optimale) signifie que les séquences de mots les plus longues auront plus d'avantage. Cela explique la détérioration de la qualité de traduction dans le cas où on augmente la taille du faisceau avec une pénalité de mots de 1 car on essaie de générer des traductions courtes avec des faisceaux plus larges.
En fixant la taille du faisceau à 20 et la pénalité de mots à 1,5, nous avons obtenu un BLEU de 24,37% sur le corpus de test.

Nous récapitulons dans le tableau 5.5, les résultats comparatifs entre l'approche statistique et l'approche neuronale.

Modèles	**BLEU**	**OOV**
Statistique	24,5%	5,7%
Neuronal	5,4%	
Neuronal+Attention+Encodage bidirectionnel	20,6%	
Neuronal+Attention+Encodage bidirectionnel+RepUnk	21,1%	
Neuronal+Attention+Encodage bidirectionnel+RepUnk+Faisceau=20+$p_w = 1,5$	24,4%	

TABLE 5.5 – Performances de traduction comparatives entre l'approche statistique à base de segments, l'approche neuronale avec/sans la technique d'attention, en traitant les mots hors vocabulaire et en optimisant la taille du faisceau ainsi que la pénalité de mots.

À partir de toutes ces expérimentations, nous arrivons à la conclusion qu'avec peu de données d'apprentissage et sur la paire de langues arabe-anglais, l'approche statistique reste meilleur que l'approche neuronale. Cette dernière nécessite des techniques plus avancées pour arriver à des résultats de traduction similaires à ceux obtenus par l'approche statistique.

5.9 Conclusion et discussion

La traduction automatique neuronale est un paradigme qui consiste à entraîner un réseau de neurones pour transformer une phrase source entière en une phrase cible. Toutes les composantes de l'approche statistique, à savoir les modèles de traduction et de langage, sont implicitement modélisés par le réseau de neurones. Pour certaines paires de langue, cette nouvelle approche a atteint des performances impressionnantes au point où il devient presque impossible de distinguer les traductions générées par l'approche neuronale et de celles générées par un être humain. En revanche, atteindre ces résultats nécessite une grande quantité de données parallèles qui sont souvent coûteuses à construire.

Dans ce chapitre, nous avons présenté une étude comparative entre l'approche statistique et l'approche neuronale dans un cadre où peu de données d'apprentissage sont disponibles. Nous avons commencé par une description de notre modèle neuronal de base où un simple encodeur est utilisé pour projeter la phrase source dans un espace multidimensionnel et la phrase cible est générée à partir de cette représentation. Ce modèle de base a été comparé à l'approche statistique classique ; les résultats ont montré que l'approche statistique est beaucoup plus performante que l'approche neuronale de base (une différence absolue de 19% dans le score BLEU). Par la suite, dans le but d'améliorer les performances de ce dernier, plusieurs techniques avancées ont été détaillées et testées. Bien que ces techniques aient permis de réduire considérablement l'écart entre l'approche statistique et l'approche neuronale (une différence absolue de 0,17% dans le BLEU), il reste encore plusieurs problèmes à résoudre, à savoir la traduction des mots rares ou inconnus et la traduction des paires de langues peu dotées en ressources. L'avantage de l'approche neuronale est l'utilisation d'un seul composant pour le processus de traduction. En revanche, nous avons montré que dans le cadre où peu de données parallèles sont disponibles, cette architecture a besoin de composants externes et de mettre en place une architecture plus aboutie pour se rapprocher des performances de l'approche statistique.

Chapitre 6

Traduction automatique du texte *code-switché*

Sommaire

Nous abordons dans ce chapitre un phénomène très répandu dans le monde arabe, le *code-switching*. Ce phénomène se produit lorsqu'un locuteur alterne entre deux ou plusieurs langues dans son discours. Communément, le *code-switching* est utilisé dans les textes informels tels que les *tweets*, les contenus en ligne et dans le langage parlé.

Dans ce chapitre, nous étudions la présence du *code-switching* dans des textes formels, à savoir les documents des institutions multilingues. Notre étude porte sur le mélange de l'arabe et de l'anglais dans un corpus parallèle extrait de documents officiels des nations unies. À partir de ces documents nous construisons un corpus parallèle *code-switché* avec deux traductions de référence, l'une en arabe pur et l'autre en anglais pur. Nous utilisons ensuite, cette ressource pour évaluer la traduction des documents *code-switchés* vers les deux langues de référence. Depuis quelques années, les données *code-switchées* sont considérées comme du bruit lors de la phase d'apprentissage et/ou de test et elles sont écartées. Nous présentons dans ce chapitre plusieurs stratégies de traduction basées sur les approches statistique et neuronale afin de mieux traduire des textes *code-switchés* [Menacer *et al.*, 2019].

6.1 *Code-switching* dans la langue arabe

Le *code-switching* ou l'alternance codique est définie comme l'utilisation de plus d'une langue par un locuteur dans un énoncé ou un discours. Ce phénomène se produit généralement dans les communautés multilingues où les locuteurs sont connus par leur capacité à changer de langues au cours de leur communication.

Le monde arabe, ayant été colonisé à travers les décennies par différents colons, a acquis de nouvelles langues. En plus de la langue locale, généralement l'arabe, chaque région s'est imprégnée de la langue du colonisateur, et a commencé à l'introduire dans le langage parlé et dans les textes informels en particulier sur les réseaux sociaux. On peut citer à titre d'exemple la présence de plusieurs mots français dans les textes informels des pays maghrébins, alors que pour les pays du golfe, on trouve plutôt des mots anglais.

Alors que le phénomène du *code-switching* est très répandu dans le langage parlé et les textes informels, on le retrouve aussi dans les textes formels mais de manière plus restreinte. Sa présence et son utilisation se résument généralement dans les procédures des institutions multilingues comme les Nations Unies (UN), l'Union Européenne (UE) et le parlement Canadien.

La présence de ce phénomène a été une contrainte à surmonter pour la communauté TALN. Même s'il y a eu plusieurs études linguistiques du *code-switching* [Poplack, 1980, Auer, 1999, Poplack et Walker, 2003, Bullock *et al.*, 2014], le traitement informatique de ce type de données reste relativement faible. Le premier cadre formel a été proposé par [Joshi, 1982] où l'auteur a étudié le *code-switching* à travers un mécanisme de commutation entre deux systèmes grammaticaux.

La plupart des études récentes se sont concentrées sur la collecte [Abidi, 2019] et l'analyse de données, par exemple la normalisation de texte [Zhang *et al.*, 2014], l'identification de la langue au niveau des mots [Schulz et Keller, 2016], l'étiquetage morpho-syntaxique [Sarkar, 2016] et la prédiction des points du changement de langue dans les textes *code-switchés* [Solorio et Liu, 2008]. Par contre, peu d'études ont été réalisées pour des tâches TALN plus avancées, à savoir la reconnaissance automatique de la parole, la modélisation du langage et la traduction automatique. Pour ces trois applications, l'absence de données *code-switchées* est un problème difficile à résoudre. [Garg *et al.*, 2017] s'attaquent à ce problème en proposant une approche pour la modélisation de langage des textes *code-switchés*. L'idée est de combiner deux modèles de langage monolingues par le biais d'un modèle probabiliste permettant l'alternance entre les deux modèles de langage. Son modèle a été utilisé utilisé pour la reconnaissance automatique de la parole *code-switchée*.

En traduction automatique, depuis quelques années, les données *code-switchées* sont considérées comme du bruit lors de la phase d'apprentissage et/ou de test [Nguyen *et al.*, 2016]. Il n'y a pas eu de réels efforts pour traduire ce type de données ; les travaux précédents se sont concentrés sur des applications spécifiques. Par exemple, [Nguyen *et al.*, 2016] ont proposé un mécanisme pour améliorer l'alignement des mots en utilisant des données *code-switchées*. [Carpuat, 2014] a mené une étude pour détecter les segments linguistiques mixtes entre le français et l'anglais dans le corpus du Canadian Hansards collecté à partir des archives officielles du parlement canadien. Un travail similaire a été réalisé par [Sinha et Thakur, 2005] pour les langues hindi-anglais. Ils ont proposé une technique basée sur l'analyse morphologique de chaque langue pour traduire des segments contenant de l'hindi et de l'anglais mélangés en hindi ou en anglais purs.

Notre objectif dans cette partie de thèse est de traduire des segments mixtes contenant de l'arabe et de l'anglais, vers l'une ou l'autre de ces deux langues. Le premier problème que nous devons surmonter est l'absence de données parallèles *code-switchées* nécessaires pour apprendre et pour évaluer nos systèmes de traduction.

6.2 Corpus parallèle *code-switché*

Dans le monde arabe, on utilise dans les conversations quotidiennes non seulement la langue arabe mais aussi d'autres langues principalement le français et l'anglais, ainsi les conversations informelles dans les réseaux sociaux et les contenus en ligne sont souvent exprimées dans plusieurs langues simultanément. Dans cette section, nous menons une étude sur la présence du *code-switching* dans les textes formels, à savoir les comptes-rendus des institutions multilingues, pour pouvoir construire un corpus parallèle *code-switché* permettant d'étudier et d'évaluer la traduction de textes bilingues *code-switchés*.

6.2.1 Structure et traitement du corpus

Notre principal objectif est de traduire des documents *code-switchés* anglais-arabe en arabe pur ou en anglais pur. Pour ce faire, il est nécessaire de disposer d'un corpus parallèle *code-switchés* avec deux traductions de référence afin d'apprendre et d'évaluer les performances du système de traduction. Pour atteindre cet objectif, nous avons choisi de faire notre étude sur les documents des nations unies. Ce choix est justifié par la présence de termes en anglais dans les documents officiels arabes.

La plupart des documents des nations unies sont publiés dans les langues officielles de cet organisme après avoir été traduits par des traducteurs spécialisés à partir du document original. Les phrases cibles de la traduction peuvent inclure des expressions (de courts segments) dans une langue différente de la langue cible. Le passage d'une langue à une autre peut se faire dans la même phrase ou entre les phrases. Dans notre cas, nous acceptons tous les cas de passage d'une langue à une autre.

Le corpus parallèle arabe-anglais extrait de documents des nations unies est celui référencé dans MultiUN [Eisele et Chen, 2010]. Après un processus de normalisation des données, nous donnons dans le tableau 6.1 des statistiques sur la taille de ce corpus. L'étape de normalisation comprend la segmentation en mots, la détermination de la bonne capitalisation des mots et le nettoyage des corpus arabe et anglais.

Langage	Nombre de phrases	Nombre de mots	Nombre de Mots uniques
Arabe	9,7M	232,7M	690k
Anglais		275,3M	388k

Table 6.1 – Informations sur le corpus parallèle arabe-anglais de MutliUN.

Notre idée est d'utiliser ce corpus pour construire une ressource parallèle où les phrases sources sont des phrases *code-switchées* arabe-anglais et les phrases cibles sont leurs traductions en arabe et en anglais. Cette ressource sera principalement utilisée pour évaluer la traduction des données *code-switchées*.

6.2.2 Sélection des phrases *code-switchées*

Le corpus parallèle de départ est composé de phrases en arabe et de leur traduction en anglais. En analysant les phrases arabes, nous avons remarqué qu'elles peuvent inclure des segments en anglais, mais aussi des segments en français, en espagnol ou dans d'autres langues. Tous ces segments sont écrits en script latin, ce qui facilite l'identification entre les segments arabes et ceux d'autres langues. Cependant, l'identification de la langue des segments écrits en script latin est difficile, elle nécessite des approches plus avancées.

À partir du corpus original, nous avons sélectionné toutes les phrases bilingues contenant des segments en arabe et en script latin. Ce processus nous a permis de diviser le corpus en trois parties, comme illustré dans la figure 6.1.

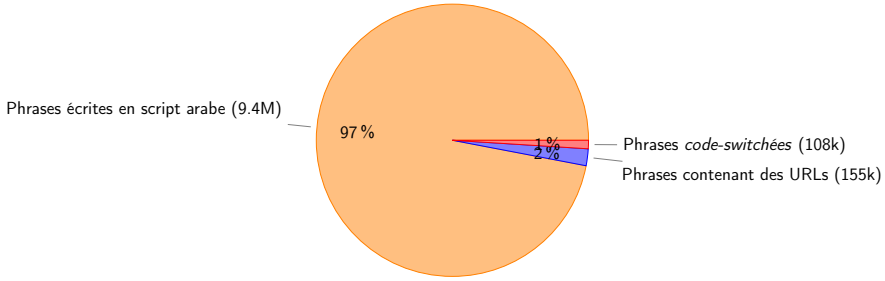


FIGURE 6.1 – Répartition des phrases arabes en fonction de la présence des segments en script latin.

Comme prévu, la majorité du texte est rédigée en arabe, 97% des phrases sont écrites en caractères arabes et seulement 3% sont mélangées. Parmi ces phrases *code-switchées*, nous n'avons gardé que celles sans liens URL, ce qui donne au total 108k phrases potentielles contenant des segments en script arabe et d'autres en script latin. Les 97% des phrases écrites en arabe pur sont utilisées plus tard pour l'apprentissage et l'évaluation de la traduction des langues pures (sans aucune instance de *code-switching*). En ce qui concerne les phrases *code-switchées*, nous avons implémenté un processus de normalisation de données comportant :

— La suppression de toutes les phrases en double.
— La suppression de toutes les phrases contenant des adresses emails.
— La suppression de toutes les phrases contenant des abréviations, par exemple : sect. pour section, p. pour page, etc.
— La suppression de toutes les phrases contenant des acronymes.

Cela réduit le nombre de phrases *code-switchées* de 108K à 55K phrases.

6.2.3 Analyse du corpus *code-switché*

Dans les 55k phrases *code-switchées* sélectionnées, l'arabe est la langue dominante ; comme illustré dans la figure 6.2, la plupart des mots sont écrits en script arabe, les mots en script latin représentent 12% des occurrences. Cela est équivalent à un nombre total de segments égal à 74k.

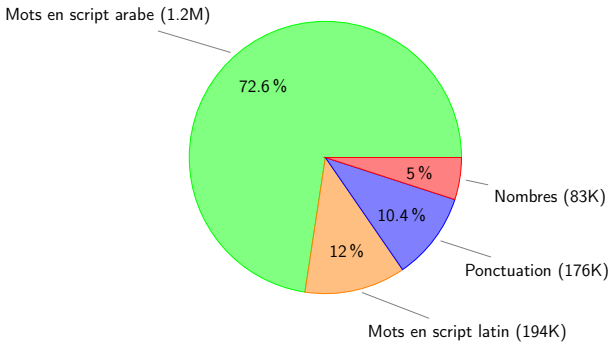


FIGURE 6.2 – Distribution des mots dans le corpus *code-switché*.

Nous donnons dans la figure 6.3 la répartition des phrases en fonction du pourcentage de

code-switching. Nous observons que 80% de phrases contiennent entre 0 et 25% de mots en script latin. Les phrases avec un pourcentage de *code-switching* compris entre 70 et 100% représentent seulement 0,3% du corpus global.

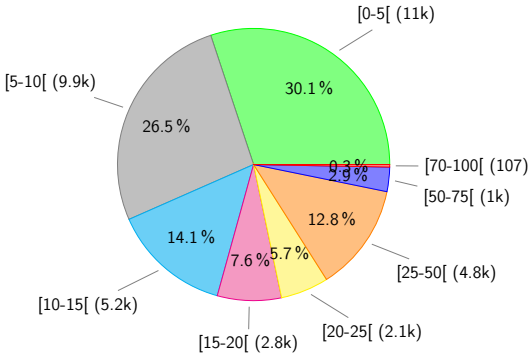


FIGURE 6.3 – Répartition du corpus *code-switché* selon le pourcentage de mots en script latin.

Bien que de nombreuses instances de *code-switching* soient dues à l'utilisation de noms ou de titres d'organisations, il existe d'autres cas où le *code-switching* est utilisé pour citer des expressions dans d'autres langues. Le tableau 6.2 présente quelques exemples de phrases *code-switchées*. Dans la plupart des cas, le *code-switching* se produit entre l'arabe et l'anglais (phrases 1 et 2 du tableau 6.2), mais d'autres langues peuvent être utilisées (le français par exemple dans la phrase 3 du tableau 6.2).

Les phrases *code-switchées*	**Traduction en français**
in due time ينبغي الَاستعَاضة عن عبَارة in a timely manner بعبَارة	L'expression "en temps voulu" doit être remplacée par "en temps opportun"
down to earth نشر رسَالة إِخبَارية بعنوَان	Diffuser un bulletin d'information intitulé pieds sur terre
fournitures courantes يشير إِلَى	Il désigne les fournitures courantes

TABLE 6.2 – Exemples de textes non arabes dans le corpus *code-switché*. Les segments en rouge représentent des segment dans des langues étrangères (anglais et français).

Afin d'étudier la répartition des langues, nous avons identifié la langue de chaque segment en utilisant *langid.py* [Lui et Baldwin, 2012], un outil d'identification des langues qui peut identifier 97 langues. Il est basé sur une classification de type naïf Bayes qui prend comme caractéristiques des n-grammes au niveau de caractères [Cavnar *et al.*, 1994]. Soixante-trois langues ont été détectées dans notre corpus *code-switché*. La distribution des langues les plus utilisées est présentée dans la figure 6.4. Comme prévu, la plupart des segments sont détectés comme des segments anglais. Il y a peu de segments provenant d'autres langues telles que le français et l'espagnol, etc.

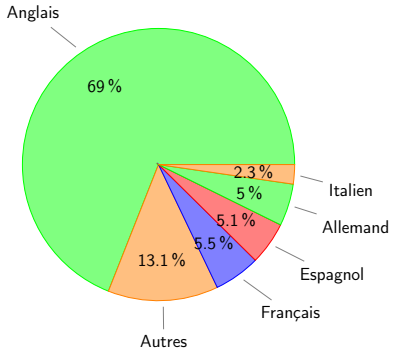


FIGURE 6.4 – Distribution des langues dans le corpus *code-switché*.

6.2.4 Construction des traductions de référence

Afin de pouvoir évaluer la traduction du corpus *code-switché* vers l'anglais et/ou l'arabe, nous avons besoin des traductions de référence dans ces deux langues. Compte tenu de la nature parallèle du corpus MultiUN, la traduction anglaise de chaque phrase *code-switchée* est facilement récupérable. Nous donnons dans le tableau 6.3 le nombre total de phrases et de mots dans le corpus parallèle *code-switché*.

Langage	**Nombre de phrases**	**Nombre de Mots**	**Nombre de Mots uniques**
code-switching	55K	1,6M	129K
Traduction de référence En		1,7M	84K

TABLE 6.3 – Informations sur le corpus parallèle *code-switché*. Les phrases sources contiennent des segments en arabe et en script latin et les phrases cibles représentent leur traduction en anglais.

Contrairement à l'ensemble des traductions anglaises de référence qui a été extrait à partir du corpus parallèle où la traduction est faite par un être humain, la traduction de référence arabe est générée par un processus automatique. L'idée principale est de traduire tous les segments en script latin dans les phrases *code-switchées* en arabe en utilisant un système de traduction automatique, et d'utiliser ensuite ces traductions pour reconstituer une nouvelle phrase avec uniquement des mots arabes. Ce processus est basé sur quatre étapes :

Extraction de segments. Cela consiste à identifier tous les segments écrits en script latin dans le corpus *code-switché*. Cette étape dépend de la détection du changement du script d'écriture.

Identification de langue. Dans notre cas, les instances de *code-switching* pourraient être entre l'arabe et l'anglais ou entre l'arabe et d'autres langues comme nous l'avons présenté dans la section 6.2.3. L'identification de la langue de chaque segment en script latin nous permet de le traduire par le système de traduction approprié.

Traduction de segments. À ce stade, les segments en script latin sont traduits en arabe pur en utilisant l'API de Google translate [17]. En raison des entités nommées, tous les segments n'ont pas été traduits en arabe. En ne gardant que les phrases *code-switchées* qui contiennent des segments traduits, le nombre total de phrases *code-switchées* est réduit de 55k à 37k.

17. https ://github.com/ssut/py-googletrans

Construction de phrases. Une fois que les segments latins sont traduits, la traduction de référence arabe est générée en remplaçant les segments non arabes par leur traduction dans les phrases *code-switchées* tout en conservant le même ordre des segments.

Nous donnons dans la figure 6.5 un exemple de génération de la traduction de référence arabe pour une phrase *code-switchée* tout en spécifiant les étapes du processus présenté ci-dessus.

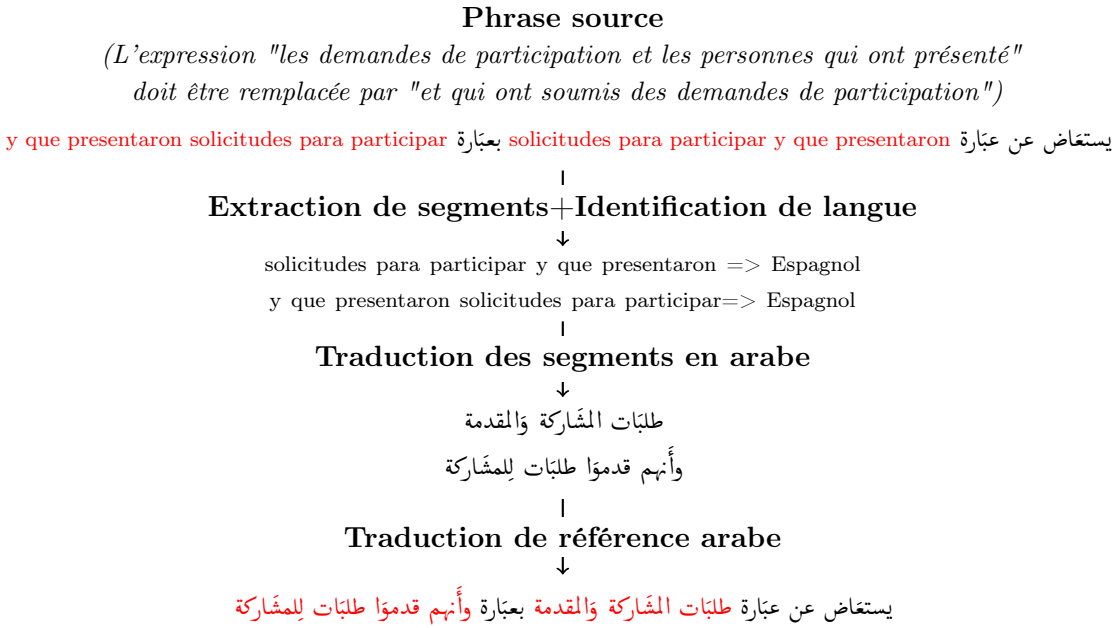


FIGURE 6.5 – Exemple de génération de la traduction de référence arabe pour une phrase *code-switchée*. Les segments en rouge représentent les segments en script latin qui sont traduits pour construire la traduction arabe finale.

6.2.5 Corpus résultant

Le tableau 6.4 présente quelques informations sur le corpus parallèle *code-switché* généré avec le processus décrit dans les sections précédentes.

Language	**Nombre de phrases**	**Nombre de mots**	**Nombre de mots uniques**
code-switching	37k	1,09M	99k
Traduction de référence En		1,14M	64k
Traduction de référence Ar		1,06M	88k

TABLE 6.4 – Informations sur le corpus parallèle *code-switché*. Ce corpus est une collection de phrases sources contenant des segments en arabe et en script latin avec leurs traductions en arabe et en anglais pur.

Quelques exemples de phrases *code-switchées* et leurs traductions de référence en anglais et en arabe sont présentés dans le tableau 6.5.

Phrase source *code-switchée*	أمّا بالَانكليزية فهي مركبة من الَاحرف الَاولَى لِلكلمَات التَالية **boosting and inspiring dynamic youth achievement**
Traduction de référence En	It is also an acronym for Boosting and Inspiring Dynamic Youth Achievement
Traduction de référence Ar	أمّا بالَانكليزية فهي مركبة من الَاحرف الَاولَى لِلكلمَات التَالية **تعزيز وَالهَام الَانجَاز الدينَاميكي لِلشبَاب**
Phrase source *code-switchée*	بالَإنكليزية **preparations for a possible transit strike** تعميم اعلَامي وَالفرنسية فقط
Traduction de référence En	Information circular , Preparations for a possible transit strike english and french only
Traduction de référence Ar	تعميم اعلَامي **الَاستعدَادَات لَاضرَاب غَابر محتمل** بالَإنكليزية وَالفرنسية فقط
Phrase source *code-switchée*	برنَامج محو الَامية المسمَى **yo si puedo**
Traduction de référence An	Program of Literacy Yes I can
Traduction de référence Ar	برنَامج محو الَامية المسمَى **نعم أستطيع**

TABLE 6.5 – Exemples de textes *code-switchés* dans le corpus parallèle *code-switché* avec leur traduction de référence en anglais et en arabe.

6.2.6 Qualité du corpus

La traduction anglaise de référence de notre corpus parallèle *code-switché* est une traduction de haute qualité générée par un être humain. Cependant, la traduction arabe de référence est un mélange de données produites par un être humain (les segments arabes dans la phrase source *code-switchée*) et de données produites par un système de traduction automatique (les segments en script latin traduites en arabe). De ce fait, et afin de pouvoir utiliser ces traductions arabes de référence pour évaluer la traduction des textes *code-switchés*, nous devons nous assurer de leur qualité ; ce qui est fait grâce à une évaluation subjective.

Pour étudier l'impact de différentes sources d'information sur la procédure d'évaluation des traductions arabes de référence, nous avons proposé deux scénarios d'évaluation différents :

Scénario 1 cible seulement : Dans ce scénario, nous n'avons affiché aux participants que la traduction arabe. Nous avons demandé à chaque participant d'évaluer le sens et la structure de la traduction générée en utilisant l'échelle à cinq points que présente le tableau 6.6. L'objectif de ce scénario est de savoir si la traduction arabe est compréhensible et respecte bien les règles de la grammaire arabe.

Scénario 2 source+cible : Dans ce scénario, nous avons affiché aux participants la phrase source *code-switchée* et sa traduction arabe. Les participants sont chargés d'évaluer la traduction en utilisant l'échelle à cinq points présentée dans le tableau 6.6. L'objectif de ce scénario est de savoir si la traduction arabe est bonne et si elle correspond bien à la phrase source.

	Scénario 1 cible seulement	**Scénario 2 source+cible**
1	Incompréhensible	Aucune relation
2	Certains segments sont compréhensibles	Certains segments sont correctement traduits
3	Compréhensible mais arabe non-natif	Traduction compréhensible mais arabe non-natif
4	Compréhensible	Traduction compréhensible
5	Excellent	Traduction parfaite

TABLE 6.6 – Échelle d'évaluation utilisée pour chaque scénario.

Comme l'évaluation subjective est coûteuse en temps, nous avons choisies aléatoirement 600 phrases pour chaque scénario, ce qui fait un nombre total de 1200 phrases à évaluer.

Dans le cadre de nos expériences, nous avons choisi six personnes dont la langue maternelle est l'arabe ; leur âge varie entre 26 et 52 ans. Puisque la plupart des instances de *code-switching* se produisent entre l'arabe et l'anglais, nous sommes assurés que le niveau d'anglais des participants était suffisant pour comprendre la phrase *code-switchée* et pour évaluer correctement la traduction.

Le score final représente la moyenne des scores attribués par les participants à chaque traduction. Les résultats sont présentés dans le tableau 6.7.

Participants	p1	p2	p3	p4	p5	p6	**Moyenne**
Scénario 1 cible seulement	2,86	3,14	3,21	3,32	3,49	3,51	3,26
Scénario 2 source+cible	4,09	4,10	4,53	4,04	3,53	4,00	4,05

TABLE 6.7 – Score moyen de l'évaluation subjective des traductions de référence arabes.

Les résultats du premier scénario d'évaluation montrent que si le participant ne lit que la traduction de référence arabe générée automatiquement sans connaître la phrase source, il trouve que la phrase est compréhensible mais avec quelques problèmes de style. Ce que nous voulons dire par un problème de style, c'est le fait que les segments ne sont pas correctement ordonnés ou qu'ils sont répétés. Cela est principalement dû à la traduction séparée des segments en script latin sans prendre en considération le contexte dans lequel ils apparaissent.
Le deuxième scénario d'évaluation montre que les participants jugent que la traduction générée est compréhensible et qu'elle correspond bien à la phrase source dans la plupart des cas.

Nous avons également mesuré la concordance entre les participants. Pour ce faire, nous avons demandé à chaque participant d'évaluer le même ensemble de phrases choisi au hasard (100 phrases). En se basant sur ces évaluations, nous avons mesuré l'accord entre les participants en calculant le coefficient de Kappa de Fleiss [Landis et Koch, 1977], une variante du kappa de Cohen qui fonctionne pour n'importe quel nombre de participants. Ce coefficient est principalement utilisé pour calculer le degré de concordance de la classification faite par les participants par rapport à ce qui pourrait être attendu si elle était faite au hasard.
Pour les deux scénarios d'évaluation, nous avons obtenu une valeur de kappa de 0,4 avec une erreur type de 0,2. Cela montre qu'il existe une concordance moyenne entre les évaluations des participants. Ce résultat était attendu car les deux classes *compréhensible/excellent* dans le scénario 1 et *traduction compréhensible/traduction parfaite* dans le scénario 2 sont si proches que le participant peut confondre entre ces deux classes.

Pour évaluer la signification statistique du kappa obtenu, nous avons calculé l'intervalle de confiance pour la valeur obtenue. Théoriquement, l'intervalle de confiance est calculé en soustrayant de la valeur de kappa la valeur du niveau de l'intervalle de confiance souhaité multipliée par l'erreur type de kappa. Dans notre cas, l'intervalle de confiance à 95% obtenu est compris entre 0,36 et 0,46. Ce résultat montre que la valeur kappa obtenue est significativement différente de zéro, ce qui confirme qu'il existe un accord moyen entre les participants.

6.3 Traduction des documents *code-switchés*

Dans cette section, nous décrivons les stratégies que nous proposons pour la traduction de textes *code-switchés*. Les stratégies proposées sont basées sur les approches statistique et neuronale, et elles sont évaluées sur le corpus parallèle *code-switché* précédemment décrit.

6.3.1 Systèmes de base

Deux systèmes de base ont été développés, un est basé sur l'approche statistique et l'autre est basé sur l'approche neuronale. Nous avons utilisé un corpus parallèle d'un million de phrases extraites du corpus MultiUN. Ce corpus est propre, c'est-à-dire que la langue source est l'arabe pur et la langue cible l'anglais pur. Il n'y a aucune instance du *code-switching* arabe-anglais dans ces données.

Le modèle fondé sur l'approche statistique utilise la même configuration que celle proposée dans le chapitre 5. Tandis que le modèle fondé sur l'approche neuronale est basé sur une architecture plus profonde que celle présentée dans le chapitre 5. Nous avons utilisé deux couches de cellules LSTM bidirectionnelles (bLSTM) pour coder la phrase source, alors que la phrase cible est décodée en utilisant deux couches de cellules LSTM. Dans le but de modéliser implicitement l'alignement entre les mots de la phrase source et ceux de la phrase cible, nous avons utilisé le modèle d'attention. L'algorithme d'optimisation *Adaptive moment estimation (Adam)* [Kingma et Ba, 2014] est utilisé pour mettre à jour et estimer les paramètres du modèle avec un *learning rate* de 2×10^{-4}. Le choix d'une architecture plus profonde est justifié par le fait qu'on dispose de plus de données d'apprentissage par rapport à l'étude menée dans le chapitre 5 (1M vs. 100k).

Lors du processus du décodage dans les systèmes de bases, les phrases *code-switchées* sont traduites entièrement sans faire de différence entre les segments arabes et ceux écrits en script latin ; ces derniers sont considérés comme des mots hors vocabulaire. Dans l'approche statistique, ces segments sont recopiés directement dans la sortie du système de traduction avec une étape de réordonnancement afin de maximiser les probabilités de l'ensemble des traductions. Dans l'approche neuronale, ces segments sont remplacés pas le mot *<unk>*.

6.3.2 Avec recopie de segments

En partant de l'idée que les segments en langue étrangère dans les phrases *code-switchées* sont bien écrits dans la langue cible, nous proposons une nouvelle stratégie de décodage. Contrairement à ce qui a été fait pour les systèmes de base où on traduisait les phrases *code-switchées* entièrement, nous allons traduire seulement les segments en script arabe alors que les segments en script latin seront directement copiés dans la sortie du système.

6.3.3 Avec augmentation du corpus d'apprentissage

Dans les deux stratégies précédentes, l'apprentissage des deux modèles (statistique et neuronal) est basé sur un corpus parallèle pur (sans aucune instance du *code-switching*) ; il serait intéressant d'utiliser un corpus parallèle *code-switché* pour apprendre les différents modèles. Le problème qui se pose est la disponibilité de cette ressource sachant que le modèle neuronal est gourmand en ressources. Notre approche pour remédier à ce problème consiste à générer ce corpus automatiquement. L'idée est de remplacer des segments arabes aléatoirement choisis dans le corpus source par leur meilleure traduction extraite à partir d'une table de traduction de 29M d'entrées. Nous avons remplacé seulement 1% de segments arabes dans le corpus source ce qui fait au total 290k segments. Nous avons choisi ce pourcentage de sorte que la distribution de segments en script latin dans le corpus résultant soit proche de la distribution de segments dans notre corpus *code-switché* de test présenté dans la section 6.2 (12%). La distribution des mots dans le corpus d'apprentissage parallèle *code-switché* généré automatiquement est présentée dans la figure 6.6.

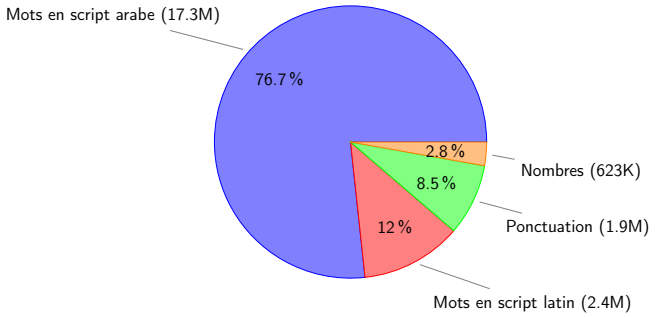


FIGURE 6.6 – Distribution des mots dans le corpus d'apprentissage parallèle *code-switché* généré automatiquement.

En se basant sur ce corpus, les mêmes techniques d'apprentissage des systèmes de base sont utilisées pour apprendre le modèle statistique et le modèle neuronal.

6.4 Résultats et discussion

Afin d'évaluer nos techniques pour traduire les textes *code-switchés*, nous avons sélectionné 5k phrases parallèles à partir de notre corpus présenté dans la section 6.2 ; nous appelons ce corpus dans ce qui suit *TestCS*. La distribution des mots dans ce corpus est présentée dans la figure 6.7.

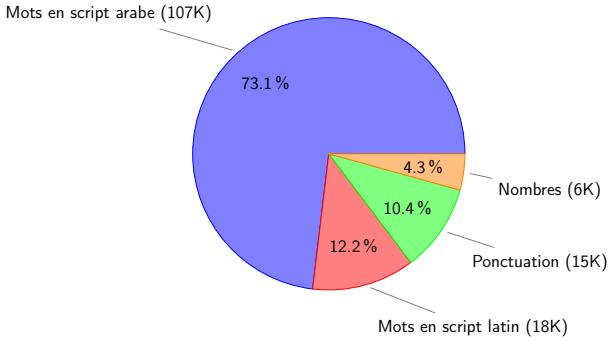


FIGURE 6.7 – Distribution des mots dans le corpus de test *code-switché* (*TestCS*).

Nous évaluons également la traduction des textes propres sans aucune instance de *code-switching*. Pour ce faire, un corpus de 5,7k phrases parallèles (*Test*), non inclus dans le corpus d'apprentissage, est utilisé. Nous donnons dans le tableau 6.8 quelques informations sur ces deux corpus de test.

Corpus	Langage	Nombre de phrases	Nombre de mots	Nombre de mots uniques
TestCS	*code-switching*	5k	146k	31k
	Traduction de référence En		152k	21k
	TTraduction de référence Ar		142k	28k
Test	Ar	5,7K	128k	18k
	En		150k	8,7k

TABLE 6.8 – Informations sur les corpus de test (le corpus *TestCS* est un corpus parallèle *code-switché* et le corpus *Test* est un corpus parallèle *propre*).

Bien que l'on dispose de la traduction de référence des phrases *code-switchées* dans les deux langues (arabe et anglais), nous avons décidé de les traduire uniquement vers l'anglais pur. Cette décision est justifiée par le fait que la plupart des segments dans les phrases *code-switchées* sont en arabe, un arabophone peut facilement comprendre le sens de la phrase selon le contexte ; il est alors intéressant de les traduire en anglais pur pour les non arabophones. Nous présentons dans le tableau 6.9 les résultats de traduction en termes de BLEU et du taux de mots hors vocabulaire *(OOV)*.

Technique	**Corpus**	**Approche statistique**	**Approche neuronale**	**OOV**
Système de base	Test	37,8	41,6	0,6
	TestCS	29,9	24,1	14,1
Avec recopie de segments	TestCS	31,1	**33,1**	14,1
Avec augmentation du corpus		**32,1**	31,1	5,7

TABLE 6.9 – Évaluation de la traduction automatique des phrases *code-switchées* (TestCS) et des phrases *propres* (Test).

Nous remarquons qu'une amélioration absolue de plus de 3,8% est obtenue avec l'approche neuronale par rapport à l'approche statistique si la traduction est effectuée de l'arabe vers l'anglais sans aucune instance du *code-switching* (corpus *Test* propre). Cela conforte notre conclusion du chapitre précédent qu'avec plus de données d'apprentissage, des techniques avancées et une architecture plus profonde, l'approche neuronale surpasse l'approche statistique. Toutefois, le plus important objectif de notre étude est la traduction du corpus *code-switché* (*TestCS*).

Dans le cas où le corpus *TestCS* est traduit en anglais pur en utilisant le système de base entraîné pour traduire de l'arabe vers l'anglais, tous les segments en script latin sont considérés comme des segments hors vocabulaire ; cela explique le taux de mots hors vocabulaire de 14,09%. En outre, nous remarquons que l'approche statistique assure de meilleurs résultats en termes de BLEU par rapport à l'approche neuronale. Cela est justifié par le processus de construction de la traduction dans chaque approche. En effet, le système statistique procède par la décomposition de la phrase source en plusieurs segments selon la table de traduction. Par la suite, tous les segments hors vocabulaire sont copiés tels quels dans la sortie. Enfin, et à cause du modèle de réordonnancement, les mots de la traduction générée sont réarrangés afin de maximiser la probabilité de traduction. En revanche, l'approche neuronale fonctionne sur un vocabulaire restreint, tous les mots qui n'existent pas dans le vocabulaire sont remplacés par le mot spécial *<unk>*, ce qui rend la traduction générée incompréhensible. Même en remplaçant les *<unk>* par les mots sources ayant générés ces derniers en se basant sur le score du modèle d'attention, cela n'améliore pas la traduction. Une explication à ces résultats serait que le modèle d'attention ne joue pas le rôle d'un alignement entre les mots de la phrase source et ceux de la cible comme dans l'approche statistique [Koehn et Knowles, 2017]. Il ne fournit qu'un modèle d'alignement simple pour aider le décodeur à décider des parties de la phrase source auxquelles il doit prêter attention [Bahdanau *et al.*, 2014].

C'est pourquoi le fait de ne pas traduire les segments en script latin et de les copier directement dans la sortie (la stratégie *avec recopie de segments* dans le tableau 6.9) améliore la traduction par rapport au système de base et en particulier dans l'approche neuronale. Pour mieux comprendre la différence entre le système de base et celui de la stratégie *avec recopie de segments*, le tableau 6.10 présente un exemple de traduction de la phrase :

"مكتب تنسيق الشُؤون الاِنسَانية ، نشرة *the humanitarian monitor* ، تموز يوليه *2009* " *(bureau de coordination humanitaire, le moniteur humanitaire, juillet 2009.)* vers la langue anglaise avec les

différents systèmes de traduction.

Phrase *CS*		2009 تموز يوليه ، the humanitarian monitor نشرة ، مكتب تنسيق الشُؤون الإنسَانية
Référence An		Office for the Coordination of humanitarian affairs , the humanitarian monitor , july 2009
Référence Ar		مكتب تنسيق الشُؤون الإنسَانية ، نشرة مرَاقب الشؤُون الإنسَانية ، تموز يوليه 2009
Système de base	**Statistique**	the office for the coordination of humanitarian affairs and *the humanitarian bulletin , monitor ,* july 2009
	Neuronale	office for the coordination of humanitarian affairs , brochure *<unk> <unk>* , july 2009 .
Avec copie de segments	**Statistique**	the office for the coordination of humanitarian affairs , bulletin *the humanitarian monitor* , july 2009
	Neuronale	the office for the coordination of humanitarian affairs , *the humanitarian monitor* , july 2009 .

TABLE 6.10 – Exemple de traduction selon les différents systèmes de traduction. Le segment en rouge dans la phrase source représentent un segment hors vocabulaire.

L'exemple du tableau 6.10 montre que la traduction de la phrase source *code-switchée* en utilisant la stratégie *avec recopie de segments* génère des traductions plus compréhensibles que ce soit avec l'approche statistique ou avec l'approche neuronale. Cela est principalement dû au fait de ne pas traduire les segments en script latin ce qui évite, d'une part, le réordonnancement des mots dans les traductions générées et d'autre part, de remplacer ces segments avec les symboles *<unk>*.

Afin de bien comprendre la relation entre les segments hors vocabulaire et la qualité de la traduction, nous avons subdivisé le corpus parallèle *code-switché* selon le pourcentage de présence du *code-switching* comme il a été présenté dans la section 6.2.3. Chaque partie du corpus global est traduite en utilisant les différents systèmes de traduction. Les résultats sont illustrés dans la figure 6.8.

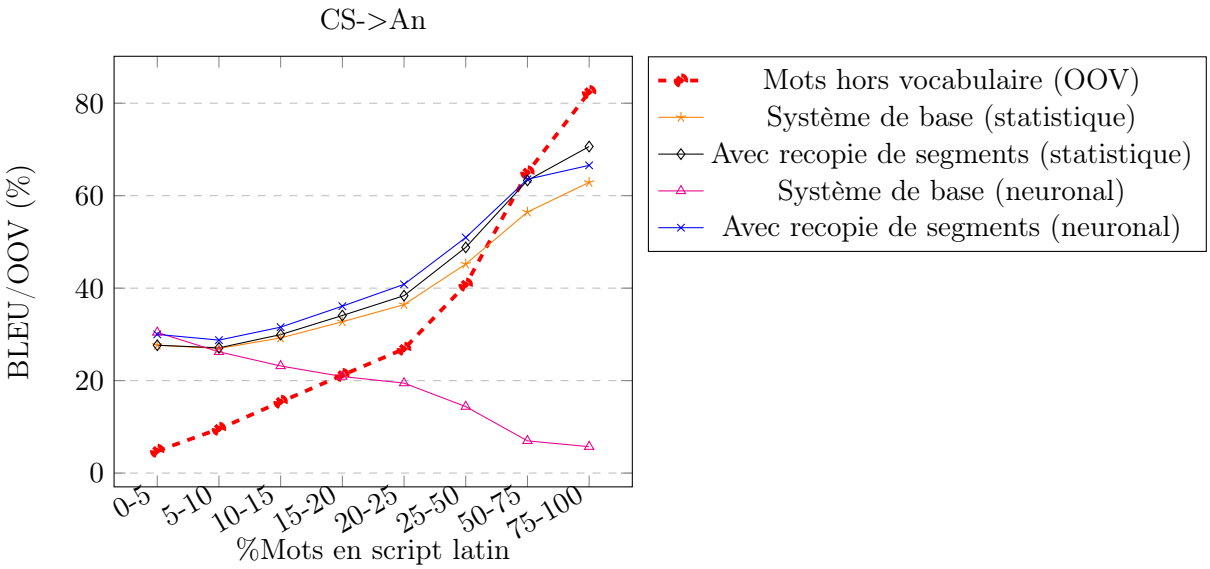


FIGURE 6.8 – Évaluation de la traduction automatique sur les sous-corpus contenant chacun un nombre homogène de segments en script latin.

Dans le cas où le pourcentage de segments en script latin est faible dans les phrases *code-switchées* (entre 0 et 5%), quelle que soit la technique de traduction, les résultats sont similaires (un BLEU entre 28 et 30%) ; cela est semblable à la traduction du texte de l'arabe vers l'anglais puisque le pourcentage du *code-switching* est très faible. Plus le pourcentage de segments en

script latin augmente, la qualité de la traduction neuronale se dégrade à cause du taux de mots hors vocabulaire élevé. En revanche, l'approche statistique assure de meilleurs résultats car ces segments en script latin sont recopiés dans la sortie du système avec une étape de réordonnancement. Encore mieux, en évitant cette étape de réordonnancement et en recopiant les segments en script latin directement dans la sortie (la stratégie *avec recopie de segments*), on arrive à améliorer la traduction de tous les systèmes proposés.

Une dernière remarque importante concernant la stratégie *avec augmentation du corpus*, un corpus artificiel *code-switché* a été généré en se basant sur une table de traduction pour apprendre les différents modèles. Bien que ce corpus ne soit pas une vraie représentation du phénomène du *code-switching*, apprendre les différents modèles sur ce dernier a donné des résultats surprenants. En effet, nous avons obtenu un BLEU meilleur que celui du système de base, mais en plus, cette approche assure de meilleurs résultats si elle est appliquée pour apprendre le modèle basé sur l'approche statistique. Ces améliorations pourraient être expliquées par la diminution du taux de mots hors vocabulaire (5,67 contre 14,09%) et par le fait que le corpus d'apprentissage se rapproche du corpus de test (*TestCS*).

6.5 Conclusion et discussion

Au cours de ces dernières années et grâce aux réseaux sociaux, l'alternance codique ou le *code-switching* est devenu plus populaire dans les communautés multilingues. Les personnes avec de fortes compétences linguistiques ont tendance à pratiquer le *code-switching* pour s'exprimer en utilisant des mots ou des expressions dans des langues étrangères. Cela s'applique très communément dans le monde arabe ; selon les régions, le vocabulaire arabe admet quelques ajouts de mots étrangers selon les contextes pour mieux exprimer des propos dans les conversations et les discussions communes ainsi que dans les réseaux sociaux.

Bien que le phénomène du *code-switching* ait une grande ampleur dans les conversations informelles, notre étude s'est basée sur sa présence dans les textes formels, à savoir les documents des institutions multilingues. L'étude réalisée dans le cadre de ce travail s'est concentrée sur l'adaptation des systèmes de traduction automatique pour traiter le *code-switching*.

Dans la communauté travaillant sur la traduction automatique, les données *code-switchées* sont généralement écartées des études et considérées comme un bruit qu'on doit éliminer. Notre premier objectif a été de proposer des stratégies pour traduire ce genre de documents. Pour ce faire, nous avons utilisé le corpus parallèle arabe-anglais extrait des documents officiels des nations unies. À partir de ce corpus, nous avons construit une ressource consistant en un texte source *code-switché* et sa traduction en arabe standard pur et en anglais pur. Cette dernière a été évaluée manuellement afin de s'assurer de sa qualité pour l'évaluation de nos stratégies de traduction des textes *code-switchés*. À notre connaissance, ce type de corpus parallèle n'existe pas ; celui que nous proposons est unique. Il pourrait également être utilisé pour des études plus approfondies sur les pratiques multilingues.

Plusieurs stratégies d'apprentissage et de traduction ont été proposées reposant sur les approches statistique et neuronale. L'apprentissage du modèle de traduction de l'approche statistique sur un corpus *code-switché* artificiel a donné des résultats de traductions surprenants. En revanche, pour l'approche neuronale, nous avons trouvé qu'en évitant la traduction des segments dans la langue étrangère, on peut avoir de bien meilleurs résultats que si on les traduit.

Chapitre 7

Projet AMIS et contributions

Sommaire

Les travaux de recherche réalisés dans le cadre de la thèse et présentés dans les chapitres précédents font partie du projet AMIS (*Access to Multilingual Information and opinionS*). AMIS est un projet européen dont l'objectif principal est d'aider les personnes à comprendre l'idée générale d'une vidéo dans une langue étrangère. Pour cela un résumé automatique de la vidéo est généré dans une langue différente de celle utilisée dans de la vidéo originale. Concrètement, répondre à ce besoin doit engager une interaction judicieuse entre plusieurs modules sujets à plusieurs défis scientifique, à savoir : l'extraction automatique d'un résumé d'une vidéo et/ou du texte, la reconnaissance automatique de la parole et la traduction automatique. Dans le cadre de cette thèse, nous nous sommes concentrés sur la reconnaissance automatique de la parole et de la traduction automatique afin de proposer un système pour traduire les vidéos.

Le projet AMIS est concerné par trois langues : l'anglais, l'arabe et le français. Nous présentons dans un premier temps le corpus collecté afin de mener les différents tests sur les composants du système final. Une approche possible pour traduire la parole consiste à appliquer successivement une étape de reconnaissance de la parole, puis une étape de traduction automatique. Le problème majeur de cette solution est la propagation des erreurs entre les deux systèmes.

Ces erreurs sont principalement dues à la disparité des domaines, *domain mismatch*, entre les données d'apprentissage des systèmes de reconnaissance automatique de la parole et le corpus du projet AMIS. Afin de minimiser les erreurs produites par le système de reconnaissance automatique de la parole, nous présentons notre approche pour adapter le vocabulaire des systèmes de reconnaissance automatique de la parole développés, à savoir : le système ALASR et deux autres systèmes pour l'anglais et pour le français [Jouvet *et al.*, 2017, Jouvet *et al.*, 2018]. Une fois que les vocabulaires sont adaptés aux données du projet AMIS, nous évaluons la qualité de la traduction automatique de vidéos en arabe vers la langue anglaise. Enfin, nous présentons une réflexion de recherche sur la possibilité d'utiliser des modèles séquence-à-séquence afin de transformer directement un signal de la parole dans une langue donnée en une séquence de mots dans une autre langue.

7.1 Projet AMIS

Avec l'émergence des réseaux sociaux et des contenus en ligne, un large flux d'actualités sous forme d'émissions directes et enregistrées sont disponibles dans différentes langues. Cette diversité de langues rend l'accès aux informations pertinentes plus difficile et l'utilisateur final n'a accès qu'à une quantité faible de ces dernières. Il s'avère que même les personnes instruites ne parlent pas plus de deux ou trois langues alors que la majorité n'en parle qu'une, ce qui explique l'inaccessibilité aux différentes informations pour la majorité des gens. Pour faciliter l'accès à ces émissions et pour aider les gens à comprendre les informations diffusées et présentées dans une langue étrangère, le projet AMIS propose de développer un système d'aide à la compréhension de l'information multilingue sans aucune intervention humaine. Le concept de compréhension est abordé par le biais du résumé automatique de vidéos afin d'en extraire les informations pertinentes et de les traduire dans une langue compréhensible par l'utilisateur. En outre, le système développé dans le cadre du projet AMIS ne permet pas seulement de visionner l'information dans sa propre langue, mais aussi de la comparer à une autre vidéo diffusée dans une autre langue et portant sur le même sujet.

Pour bien comprendre les deux objectifs du projet AMIS, considérons le scénario suivant : un voyageur dans une ville à l'étranger, il allume la télévision et un événement local survient, qui fait la une des journaux télévisés. Le voyageur zappe d'une chaîne à une autre, mais ne comprend que partiellement ce qui se passe. Pourtant, cela peut avoir un impact fort sur son séjour (fermeture d'aéroport, limitation de circulation, etc.). Le voyageur a donc besoin, depuis son hôtel, d'avoir un compte rendu rapide sur la situation dans sa propre langue. Le système AMIS lui permet d'avoir un résumé de cette vidéo dans sa propre langue et ainsi il pourra facilement comprendre de quoi il s'agit. Lorsque ce même voyageur rentre de son voyage, il peut s'apercevoir que ce même évènement est diffusé sur les chaînes locales mais présenté différemment. Cette différence est principalement due au fait que l'évènement a été présenté en langue étrangère où plusieurs facteurs ont un impact sur la manière de présenter les évènements (par exemple la culture, la politique, la religion, etc.). Le système AMIS propose également une comparaison de l'événement présenté dans deux langues en termes de divergence, convergence et émotions présentes dans les deux vidéos.

Concrètement, répondre à ce besoin implique plusieurs défis scientifiques : extraire un résumé d'une vidéo en se basant sur l'information vidéo (résumé de vidéo) et sonore (résumé de texte), traduire le flux sonore dans une autre langue (traduction parole-texte, ou même parole-parole) et analyser les sentiments pour comparer deux vidéos présentées dans deux langues différentes. De ce fait, le système développé dans le cadre du projet est basé sur l'utilisation de plusieurs com-

posantes nécessitant une interaction judicieuse entre les différents modules [Smaïli *et al.*, 2018]. Les architectures développées dans le cadre de ce projet font appel aux composantes suivantes : résumé automatique de vidéos/textes, reconnaissance automatique de la parole, segmenteur de transcription, traduction automatique et analyse de sentiments.

7.2 Corpus AMIS

Dans le cadre du projet AMIS, nous nous intéresserons à trois langues : la langue arabe et deux langues européennes, l'anglais et le français. L'arabe a été sélectionnée car, d'une part, elle est complètement différente des langues anglaise et française, et d'autre part, la comparaison serait intéressante entre une vidéo en anglais et une autre en arabe vu que la culture, la politique et la religion dans le monde arabe sont différentes de celles du monde occidental. L'un des premiers défis dans le cadre du projet était la collecte de diverses vidéos qui portent sur des sujets similaires dans les trois langues. [Koźbiał et Leszczuk, 2019] ont choisi YouTube comme source de vidéos où la sélection était basée sur une liste de *hashtags* portant sur des sujets polémiques. En outre, une liste de chaînes a été sélectionnée en s'assurant que celles-ci fournissent des vidéos sous forme de bulletins d'information ou de reportages.

Le corpus final contient des milliers de vidéos englobant 100 heures d'enregistrements pour chaque langue comme illustré dans le tableau 7.1. Ce corpus est principalement utilisé pour les tests finaux du système développé dans le cadre du projet AMIS.

Langue	**Anglais**	**Arabe**	**Français**
Nombre de vidéos	1874	1503	2046

TABLE 7.1 – Nombre de vidéos par langue.

Les corpus que nous avons utilisés dans le chapitre 3 pour le développement de notre système ALASR ont été collectés entre la période de novembre 2001 et décembre 2010. Or le corpus de test d'AMIS est une collection de vidéos plus récentes. Cela signifie que le vocabulaire du système de reconnaissance automatique de la parole a été défini en fonction des corpus de textes qui sont beaucoup plus anciens. Par conséquent, le vocabulaire ne reflète pas correctement les noms propres observés dans les vidéos récemment collectées du projet AMIS. La même remarque s'applique aux systèmes de reconnaissance pour les langues anglaise et française développés dans le cadre du projet AMIS. Dans ce qui suit, nous présentons brièvement les systèmes de reconnaissance automatique de la parole de base pour chacune des langues. Nous présentons ensuite notre approche pour récupérer la transcription de référence de quelques vidéos du corpus AMIS afin d'évaluer les performances des systèmes de reconnaissance de base.

7.3 Systèmes de base

Le développement des systèmes de base pour les langues anglaise, arabe et française est basé sur la même *recette*, à savoir celle utilisée dans le système ALASR (voir chapitre 3). Rappelons que le modèle acoustique est à base de réseau de neurones de type perceptron multicouches. Pour la modélisation du langage, des modèles n-grammes classiques ont été utilisés. Nous donnons dans le tableau 7.2 des informations sur les ressources utilisées pour développer les systèmes de reconnaissance automatique de la parole pour chaque langue.

Langues	Nombre de mots (données textuelles)	Taille du vocabulaire	Prononciations/mot	Donnée orales
Anglais	155M	150k	1,1	200h
Arabe	1M	95k	5,1	52h
Français	1,6M	97k	2,1	200h

TABLE 7.2 – Informations sur les données textuelles et orales utilisées pour le développement ses systèmes de reconnaissance automatique de la parole pour les trois langues.

Le nombre de prononciations par mot dans le lexique de l'arabe est élevé à cause de l'absence des diacritiques dans le texte. En français, les variantes de prononciation sont dues au *e* muet à la fin de nombreux mots, et aux consonnes de liaison avec les mots suivants commençant par une voyelle. Pour les trois langues, le taux d'erreur mot varie entre 13% et 14%. En revanche, en évaluant ces systèmes sur le corpus AMIS, nous avons obtenu un taux d'erreur plus élevé. Dans ce qui suit, nous présentons notre démarche pour faire cette évaluation sachant qu'aucune transcription n'est disponible pour les vidéos du corpus AMIS.

7.4 Évaluation sur le corpus AMIS

Comme nous l'avons mentionné précédemment, notre corpus de test est une collection de vidéos YouTube. En analysant les chaînes à partir desquelles ces vidéos ont été collectées, nous avons trouvé une description associée à certaines vidéos. Ces descriptions sont principalement relatives aux vidéos de la chaîne Euronews. Elles sont accessibles directement à partir de YouTube ou du site officiel d'Euronews. La figure 7.1 illustre une page web qui récapitule un exemple de description d'une vidéo en français. On peut remarquer que la description de YouTube est courte (seulement quatre lignes pour cette vidéo) tandis que la description d'Euronews est plus longue.

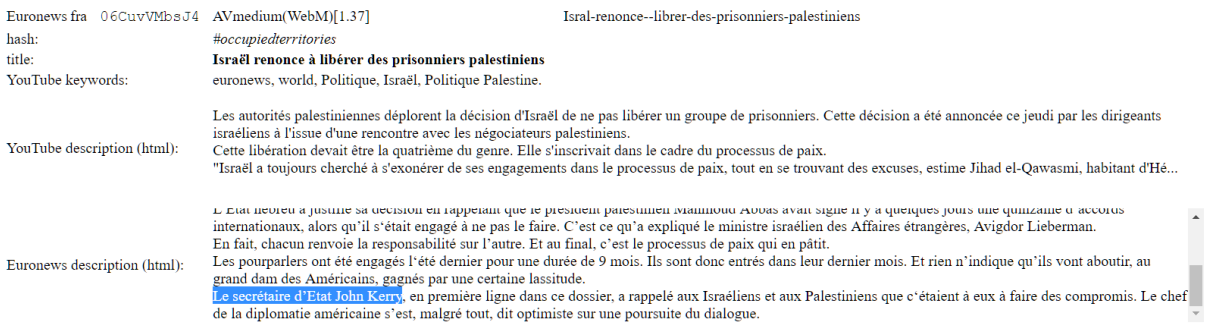


FIGURE 7.1 – Exemple de descriptions de YouTube et d'Euronews pour une vidéo en français. En plus de la description, le titre, les mots clé et le *hashtag* utilisé pour collecter la vidéo sont affichés.

Dans la plupart des cas, ces descriptions correspondent approximativement à la transcription des vidéos. De ce fait, nous avons décidé d'en extraire quelques phrases de référence et de les utiliser comme une transcription de référence *approximative* afin d'évaluer les performances des systèmes de base.

7.4.1 Extraction des transcriptions de référence

En analysant les descriptions provenant de YouTube ou d'Euronews, nous avons pu constater que de nombreuses séquences de mots correspondent bien à ce qui a été prononcé dans la vidéo. En contrepartie, d'autres séquences ne correspondent pas ou elles ne sont pas au bon endroit. Notre objectif est donc d'extraire des phrases entières correspondant au mieux au signal de parole et de les utiliser ensuite comme des transcriptions *approximatives* pour l'évaluation des performances des systèmes de base de reconnaissance automatique de la parole. Notre idée est d'aligner les descriptions avec la sortie des systèmes de reconnaissance automatique de la parole. Pour accomplir cette tâche, nous avons optimisé plusieurs scores calculés en se basant sur les séquences de mots de la description et celles prédites par les systèmes de reconnaissance. Ces scores sont :

— Le score de substitution entre les mots. Il est principalement basé sur le calcul de la différence en termes de caractères entre ces mots, ce qui signifie qu'un score plus faible est attribué aux mots similaires.
— Les scores d'insertion et de suppression qui dépendent de la longueur des mots. Un faible score est attribué aux mots courts alors qu'un score plus élevé est donné aux mots plus longs.

Vu que la segmentation des descriptions de vidéos en phrases est primordiale pour pouvoir les comparer avec la sortie des systèmes de reconnaissance automatique de la parole, nous nous sommes basés sur les ponctuations pour détecter le début et la fin des phrases. Pour sélectionner ensuite les phrases de référence, nous avons accordé une plus grande attention aux coûts de correspondance au début et à la fin des phrases. En effet, les phrases pour lesquelles le coût d'alignement sur les trois premiers mots et sur les trois derniers mots est inférieur à un seuil donné sont sélectionnées comme phrases de référence. Il est à noter que toutes les phrases dont le taux d'insertion et de suppression est supérieur à 30% ne sont pas sélectionnées comme phrases de référence.

Nous donnons dans le tableau 7.3 des informations sur les transcriptions de référence *approximatives* récupérées suite à l'alignement des descriptions de YouTube et d'Euronews avec la sortie des systèmes de reconnaissance automatique de la parole.

Langues	**Description YouTube**		**Description Euronews**	
	Nombre de phrases	**Nombre de mots**	**Nombre de phrases**	**Nombre de mots**
Anglais	2,6k	51k	4,9k	77k
Arabe	0,8k	21k	0,5k	11k
Français	5,6k	119k	7,3k	146k

TABLE 7.3 – Informations sur les transcriptions de référence *approximatives* des trois langues.

Le tableau 7.3 montre que l'on a obtenu plus de données de référence pour le français que pour l'anglais ou l'arabe. En effet, les traitements précédents ont permis d'extraire près de 50% du texte des descriptions comme données de référence pour cette langue. Tandis que ce chiffre baisse à 30% pour l'anglais et pour l'arabe. Un autre point intéressant qui peut être observé est la longueur des phrases de référence. Pour les trois langues et pour les deux sources d'information (YouTube et Euronews), les phrases extraites ont une longueur assez similaire, en moyenne 20 mots par phrase.

7.4.2 Évaluation des systèmes de reconnaissance automatique de la parole

Les transcriptions *approximatives* récupérées à partir du YouTube et d'Euronews sont utilisées pour avoir une estimation du WER pour le corpus AMIS. Le tableau 7.4 illustre l'évaluation en termes du WER des résultats de reconnaissance automatique de la parole sur le corpus AMIS.

Langues	**WER** (Corpus Test)	**WER** (Corpus AMIS-YouTube)	**WER** (Corpus AMIS-Euronews)
Anglais	13,0%	17,8%	17,2%
Arabe	14,4%	27,0%	27,2%
Français	14,2%	17,4%	17,8%

TABLE 7.4 – Taux d'erreur mot obtenus sur les corpus de test (utilisés pour évaluer les systèmes de reconnaissance automatique de la parole de base) et sur le corpus AMIS.

Nous pouvons remarquer que les estimations des taux d'erreur sur le corpus AMIS sont plus élevées par rapport à celles obtenues sur les corpus de test. Ces résultats étaient attendus car le texte utilisé comme référence pour les vidéos du corpus AMIS n'est pas la transcription exacte de ce qui a été prononcé dans les vidéos. Pour la langue française, le taux d'erreur est de 17,4% sur les références provenant du YouTube, et de 17,8% sur les références d'Euronews. Dans les deux cas, le pourcentage de suppression est de 3,5% et d'insertion est de 4%. Sur les données anglaises, nous avons obtenu des taux d'erreur similaires à ceux de la langue française : 17,8% sur les références du YouTube, et 17,2% sur les références d'Euronews. Pour les taux de suppression et d'insertion, ils sont similaires à ceux du français (environ 3,5 à 4,2%). En ce qui concerne les données de l'arabe, nous avons obtenu un taux d'erreur plus élevé. Cela est principalement dû à la manière d'orthographier les mots en arabe. Les erreurs les plus fréquentes rencontrées dans les résultats sont la substitution des mots avec la lettre *Alif* ا (إتحَاد et اتحَاد (union)) et l'orthographe des noms propres (سوريَا et سورية (Syrie)).

Pour une analyse plus profonde des taux d'erreur obtenus sur le corpus AMIS, nous avons analysé grammaticalement la sortie du système de reconnaissance automatique de la parole de la langue française afin de calculer le taux d'erreur sur les entités nommées (noms de personnes, d'endroits, etc.). La raison derrière le choix des entités nommées est que cette catégorie est la plus évolutive au cours du temps ; il est donc intéressant de connaître les noms propres qui sont mal reconnus par le système de base. Pour ce faire, nous avons annoté la transcription *approximative* des vidéos en français en terme de catégories grammaticales en utilisant Treetagger [Schmid, 1994]. Nous avons trouvé un taux d'erreur de 39% sur les entité nommées contrairement au taux d'erreur global qui était moins de 18%. Nous avons aussi trouvé que 14% de ces mots sont des mots hors vocabulaire. Afin de couvrir ces mots et d'avoir des pourcentages qui se rapprochent de ceux obtenus sur les corpus des systèmes de base, nous présentons dans la section suivante les données utilisées pour l'adaptation des vocabulaires des systèmes de reconnaissance automatique de la parole.

7.5 Données textuelles pour l'adaptation du vocabulaire

L'analyse élaborée dans la section précédente montre que les vocabulaires des systèmes de base de reconnaissance automatique de la parole ne sont pas à jour et par conséquent les noms propres observés dans le corpus AMIS y sont absents. Rappelons que les corpus utilisés dans les systèmes de base ont été collecté entre les années 90 et 2000, contrairement au corpus AMIS qui

est récent. Pour adapter les vocabulaires aux vidéos collectées dans le cadre du projet AMIS, de nouvelles données textuelles ont été collectées correspondant à la période de production de ces vidéos. Nous décrivons dans ce qui suit ces données textuelles.

7.5.1 Données d'apprentissage

Pour avoir un maximum de diversité dans les données textuelles, nous avons collecté sur plusieurs mois des données textuelles à partir de plusieurs sites web. Ces sites sont principalement ceux de journaux, de stations radio ainsi que de chaînes de télévision dans les trois langues. Le tableau 7.5 donne quelques informations sur les données collectées. Ces chiffres sont calculés après une phase de normalisation de données qui consiste principalement à supprimer du texte inutile comme les liens URL, les adresses emails, le texte en script latin dans les documents arabe, etc. Par ailleurs, nous avons supprimé toutes les phrases en double. Avec ces traitements, près de 80% des données collectées sont ignorées.

Langues	**Données du web**	**Gigaword**
Anglais	2,9G	4,1G
Arabe	0,7G	1,1G
Français	1,9G	0,8G

TABLE 7.5 – Nombre de mots dans les données textuelles collectées et dans les corpus Gigaword utilisés dans les systèmes de base.

Nous avons calculé le pourcentage de couverture des mots les plus fréquents dans les données collectées et dans les données utilisées dans les systèmes de base de reconnaissance automatique de la parole à savoir le corpus Gigaword pour chaque langue (voir tableau 7.2). Les résultats de cette analyse sont illustrés dans la figure 7.2. Les courbes en pointillés représentent le pourcentage de couverture des mots les plus fréquents dans les corpus Gigaword, tandis que les courbes continues correspondent aux corpus collectés à partir du web.

Pour chacune des langues, les résultats de la figure 7.2 montrent que l'évolution des pourcentages de couverture est la même sur les deux types de corpus (Gigaword et du web). Nous avons trouvé que pour l'anglais, 96,6% de données textuelles collectées (2,9G mots) sont couvertes par les 100k mots les plus fréquents (la courbe continue rouge). Le même pourcentage est observé pour la langue française. En revanche, pour la langue arabe, nous avons trouvé que les 100k mots les plus fréquents couvrent 92,7% de données collectées sur le web (la courbe continue bleue). Ce faible pourcentage par rapport à celui observé pour l'anglais et pour le français s'explique par la richesse morphologique de la langue arabe.

7.5.2 Données de validation et de test

Pour la validation de notre approche d'adaptation des vocabulaires, nous avons collecté d'autres données textuelles à partir du site d'Euronews sur la même période que pour le corpus AMIS. Elles correspondent à la description [18] de 8 000 vidéos en arabe et leurs équivalentes en anglais et en français.

Par ailleurs, l'évaluation de notre approche d'adaptation des vocabulaires est basée sur les descriptions collectées pour le corpus AMIS (voir section 7.4.1). Nous donnons dans le tableau 7.7 le nombre d'occurrence de mots dans les corpus de validation et de test.

18. Rappelons que la description correspond au texte associé à une vidéo données.

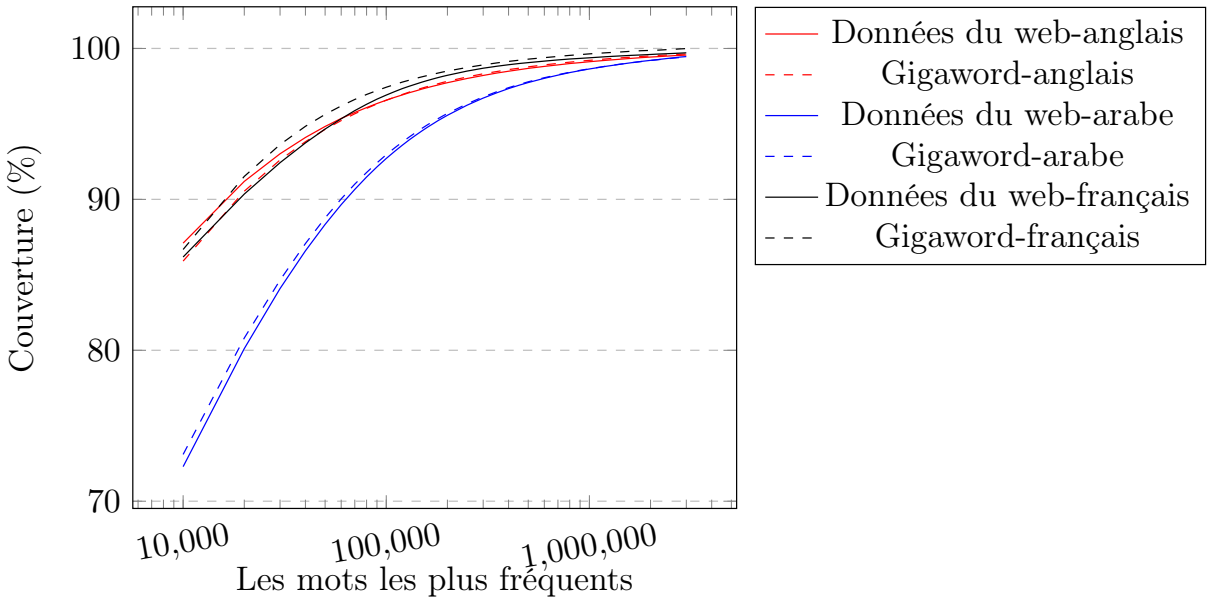


FIGURE 7.2 – Évolution du pourcentage de couverture des mots les plus fréquents dans les corpus textuels (Gigaword et du web).

Langues	**Validation**	**Test**
Anglais	1 720k	280k
Arabe	1 240k	70k
Français	1 500k	250k

TABLE 7.6 – Nombre d'occurrence de mots dans les corpus de validation et de test.

7.6 Adaptation du vocabulaire des systèmes de reconnaissance

Les données collectées à partir du web et présentées dans la section précédente sont utilisées pour sélectionner de nouveaux vocabulaires dans le but de remplacer ceux des systèmes de base et pour transcrire par la suite les vidéos du corpus AMIS. L'évaluation de notre approche est basée sur l'analyse des taux de mots hors vocabulaire sur le corpus AMIS en fonction de la taille des nouveaux vocabulaires dans les trois langues. Pour la langue française, une analyse plus détaillée sur les performances du système de reconnaissance automatique de la parole est établie.

7.6.1 Sélection du vocabulaire

Le processus de sélection des vocabulaires est basé sur une approche classique où les mots les plus probables dans la collection de données du web sont sélectionnés pour qu'ils soient intégrés dans les nouveaux vocabulaires. Les probabilités utilisées pour la sélection des mots sont celles estimées par un modèle de langage unigramme entraîné sur nos données collectées à partir du web. Vu que ces données proviennent de plusieurs sources d'information, nous les avons regroupées en plusieurs sous-groupes en fonction de la source d'information (journaux, chaînes de télévision, chaînes radio, etc.). Pour le français, nous avons défini 30 sous-groupes correspondant par exemple aux données provenant d'Euronews, de France24 et des journaux : Le figaro, Le monde, etc. Nous avons défini 22 sous-groupes pour l'anglais et 29 sous-groupes pour l'arabe. La

première étape de traitement consiste à estimer un modèle unigramme sur chaque sous-groupe de données.

Les modèles unigrammes obtenus ont été interpolés linéairement pour concevoir le modèle unigramme final. Les poids d'interpolation ont été déterminés sur le corpus de validation de sorte que la perplexité sur ce corpus soit minimale. Pour les trois langues, les plus grands coefficients d'interpolation sont ceux associés au sous-groupe de données provenant d'Euronews (des valeurs supérieures à 0,8). Ces résultats s'expliquent par le fait que le corpus de validation, sur lequel les poids sont optimisés, provient également d'Euronews.

Une fois le modèle unigramme obtenu à partir des données collectées, les vocabulaires sélectionnés correspondent aux mots les plus probables dans le modèle unigramme interpolé. Pour chaque langue, quatre vocabulaires ont été extraits en fonction du nombre de mots ; ils correspondent respectivement aux 100k, 200k, 400k et 800k mots les plus probables.

7.7 Résultats et discussion

Afin d'étudier l'impact des nouveaux vocabulaires, nous avons calculé le taux de mots hors vocabulaire sur les transcriptions de référence *approximatives* du corpus AMIS récupérées à partir des descriptions de YouTube (voir section 7.4.1). Les résultats obtenus pour les trois langues avant et après l'adaptation des vocabulaires sont illustrés dans le tableau 7.7. Nous donnons aussi les pourcentages des mots hors vocabulaire sur l'ensemble de validation déjà présenté dans la section 7.5.2.

(a) Données de validation

	Anglais	**Arabe**	**Français**
Nb mots	1 720k	1 240k	1 500k
Nb mots uniques	64k	129k	51k
Vocab de base	7,2%	17,4%	1,8%
Vocab 100k	1,1%	5,5%	0,4%
Vocab 200k	0,4%	3,1%	0,1%
Vocab 400k	0,3%	1,5%	0,1%
Vocab 800k	0,3%	0,2%	0,1%

(b) Données de test

	Anglais	**Arabe**	**Français**
Nb mots	280k	70k	250k
Nb mots uniques	21k	20k	20k
Vocab de base	5,5%	16,4%	1,8%
Vocab 100k	3,3%	6,8%	0,8%
Vocab 200k	2,7%	4,5%	0,4%
Vocab 400k	1,9%	3,1%	0,2%
Vocab 800k	1,5%	2,0%	0,2%

TABLE 7.7 – Taux de mots hors vocabulaire avant et après l'adaptation des vocabulaires.

Nous pouvons remarquer qu'en utilisant des vocabulaires de même taille que ceux des systèmes de base (100k mots), nous obtenons des taux de mots hors vocabulaire plus faibles ; nous avons atteint des différences absolues de 2,2%, 9,6% et 1% sur les données de test pour l'anglais, l'arabe et le français respectivement. Nous avons remarqué le même comportement sur les données de validation pour les trois langues. En augmentant la taille des nouveaux vocabulaires, on arrive à réduire considérablement les taux de mots hors vocabulaire.
Les taux de mots hors vocabulaire sur les données de la langue française sont plus faibles que ceux observés sur la langue anglaise, tandis que les taux les plus élevés sont ceux observés pour l'arabe. Ces résultats confirment le constat du chapitre 3 qui concerne le taux de mots hors vocabulaire dans la langue arabe : pour le même nombre d'entrée dans le vocabulaire, le taux de mots hors vocabulaire en arabe est toujours plus élevé par rapport à l'anglais.

7.7.1 Reconnaissance automatique de la parole pour le français

Afin d'étudier l'apport des nouveaux vocabulaires sur les performances du système de reconnaissance automatique de la parole, nous avons généré automatiquement la transcription de

notre corpus de test avec le système de base et avec le nouveau vocabulaire de 100k mots. La figure 7.3 illustre un exemple de transcription d'une vidéo en français générée par le système de reconnaissance automatique de la parole avant et après l'adaptation du vocabulaire. Cet exemple montre l'utilité de l'adaptation des vocabulaires pour les données de test en particulier en ce qui concerne la reconnaissance des noms propres.

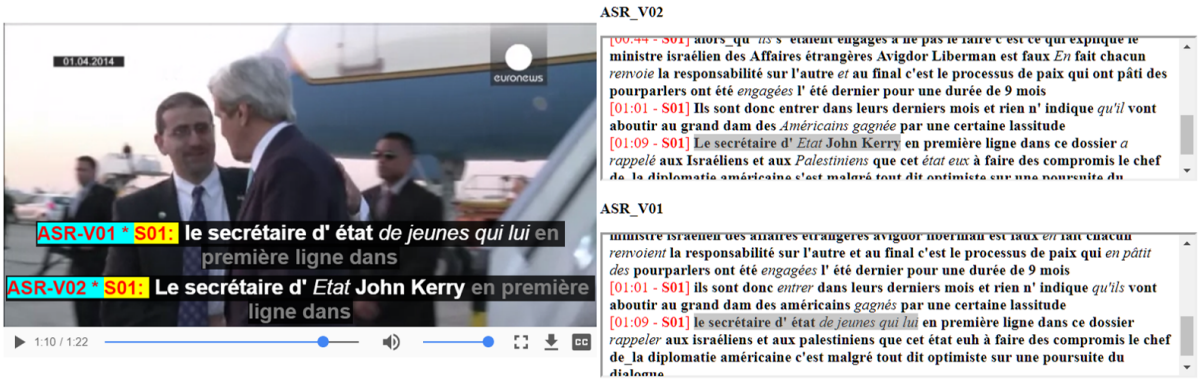


FIGURE 7.3 – Exemple de la sortie du système de reconnaissance automatique de la parole pour une vidéo en français en utilisant le vocabulaire de base (ASR_V01) et celui adapté pour le corpus AMIS (ASR_V02).

Dans l'exemple de la figure 7.3, le nom *John Kerry* n'a pas été reconnu correctement par le système de base ; il a été remplacé par une séquence de mot qui est proche : *de jeunes qui lui*, car le nom *John Kerry* est hors vocabulaire. Ce phénomène rend d'une part la sortie du système de reconnaissance automatique de la parole incompréhensible, et cela impacte d'autre part, le processus de traduction automatique.

Les résultat de reconnaissance automatique de la parole en termes du WER sont présentés dans le tableau 7.8.

	Vocabulaire de base	**Vocabulaire 100k**
WER	17,4%	14,9%
Pourcentage de mots hors vocabulaire	1,3%	0,8%
WER sur le noms propres	40%	27%
Pourcentage de mots hors vocabulaire sur le noms propres	12%	5%

TABLE 7.8 – Performances du système de reconnaissance automatique de la parole sur un corpus de test français avant et après l'adaptation du vocabulaire.

Les tests du tableau 7.8 ont été réalisés sur un ensemble réduit de 6,9k mots extraits à partir du corpus de test AMIS (voir section 7.5.2). Nous remarquons qu'en utilisant le nouveau vocabulaire de 100k mots, le taux d'erreur a baissé de 2,5%. Cela est, entre autres, dû à la réduction du taux de mots hors vocabulaire (1,3% v.s. 0,8%). En calculant les taux d'erreur sur les noms propres, nous avons trouvé une différence absolue de 13% entre le système de base et le système utilisant le nouveau vocabulaire.

7.8 Traduction automatique de la parole

7.8.1 Système séquentiel

L'un des principaux objectifs du projet AMIS est de traduire une vidéo donnée (principalement en arabe ou en français) vers une autre langue (principalement l'anglais). Une approche de base consiste à utiliser une architecture séquentielle où la sortie du système de reconnaissance automatique de la parole est utilisée comme entrée du système de traduction automatique comme illustré dans la figure 7.4.

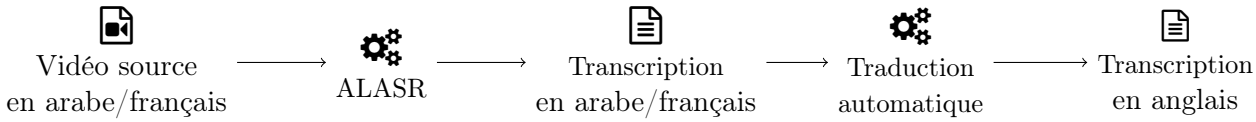


FIGURE 7.4 – Architecture séquentielle pour la traduction automatique de la parole.

Afin d'évaluer les traductions générées par le système séquentielle, nous avons utilisé comme corpus de test une partie des transcriptions *approximatives* du corpus AMIS collectées à partir de YouTube et d'Euronews (voir Section 7.4.1). Ce corpus comporte 197 vidéos en arabe correspondant à 1 253 phrases. Nous avons évalué principalement la traduction de l'arabe vers l'anglais. Pour ce faire, on a eu besoin d'une traduction de référence pour ces transcriptions en anglais, or cette dernière n'était pas disponible. Pour y remédier, nous avons décidé d'en construire une en utilisant l'API de Google translate [19]. Ensuite, la traduction générée par notre système basé sur l'approche statistique (voir chapitre 5) est évaluée en termes de BLEU en considérant comme traduction de référence celle générée automatiquement par l'API de Google translate. Nous avons obtenu un BLEU égal à 26,7% ce qui est considéré comme très acceptable.

Le problème majeur de cette architecture est l'accumulation des erreurs. En effet, nos systèmes de reconnaissance automatique de la parole produisent des taux d'erreurs variant entre 14 et 17% sur notre corpus de test AMIS. Ces erreurs sont transmises aux systèmes de traduction automatique ce qui génère par conséquent un texte entaché d'erreurs.

7.8.2 Système de bout en bout pour la traduction de la parole

Une approche alternative pour remédier au problème que pose le modèle séquentiel est de proposer une architecture de bout en bout, *end2end*, comme celle utilisée dans les approches neuronales de la traduction automatique de texte. L'idée de cette approche est d'utiliser les modèles séquence-à-séquence déjà présentés dans le chapitre 1 section 1.2.2 pour traduire de la parole. Initialement, ces modèles ont été proposés pour la traduction automatique [Sutskever *et al.*, 2014] et adaptés par la suite pour la reconnaissance automatique de la parole [Chan *et al.*, 2015]. Ce qui fait la différence entre l'application de ces modèles dans la reconnaissance automatique de la parole et dans notre tâche (la traduction automatique de la parole) ce sont les entrées et les sorties du modèle. Rappelons que les modèles séquence-à-séquence prennent en entrée des observations acoustiques. Un encodeur projette, ensuite, les vecteurs acoustiques dans un espace multidimensionnel pour générer des vecteurs de contexte. Enfin, le décodeur génère les unités acoustiques qui correspondent au mieux au signal d'entrée en se basant sur les vecteurs de contexte. Dans la tâche de reconnaissance automatique de la parole, le signal d'entrée et les unités acoustiques correspondent à la même langue. Dans le cas de la traduction automatique

19. https ://github.com/ssut/py-googletrans

de la parole, le signal d'entrée est dans une langue et les unités acoustiques correspondent à une autre langue. La difficulté majeure à ce niveau est de trouver des corpus d'apprentissage alignés au niveau des phrases de la forme : [signal acoustique dans une langue A, transcription dans une langue B]. Notre première étape consiste à construire une telle ressource (appelé plus loin corpus multilingue) afin d'apprendre et de tester un modèle séquence-à-séquence qui génère la transcription d'une vidéo dans une langue différente de celle utilisée dans la vidéo originale.

Corpus multilingue

Notre objectif est de collecter un corpus où le signal vocal est dans la langue arabe et la transcription est dans la langue anglaise. Malheureusement, nous n'avons pas trouvé un tel corpus. Pour cette raison, nous avons décidé de chercher un corpus avec un signal vocal en anglais et une transcription en arabe/français. Pour ce faire, nous avons exploité le contenu en ligne des conférences TED [20]. L'avantage des conférences TED est qu'elles sont traduites en plusieurs langues par des êtres humains. Un exemple de transcription en français pour une vidéo en anglais est donné dans la figure 7.5.

FIGURE 7.5 – Exemple d'une transcription en français où la vidéo originale est en anglais. Pour cette vidéo, la transcription est disponible en deux langues anglais et français. La transcription est divisée en segments tout en indiquant la durée début et fin de chaque segment. La traduction est faite par un humain.

Nous avons utilisé un script pour explorer le contenu du site TED correspondant à plusieurs mois. Nous avons collecté au total 1 342 vidéos en anglais ce qui est équivalent à 313 heures. Pour chaque vidéo, on peut trouver plusieurs transcriptions en différentes langues comme le montre l'histogramme de la figure 7.6 ; nous donnons le nombre de vidéos ayant été traduites dans d'autres langues selon leur fréquence.

Parmi les 1 342 vidéos en anglais (313 heures), nous avons pu collecter la transcription de 1 320 vidéos (308 heures) en espagnol. Mais, ce qui nous intéresse le plus est le nombre de

20. https://www.ted.com/talks?language=fr

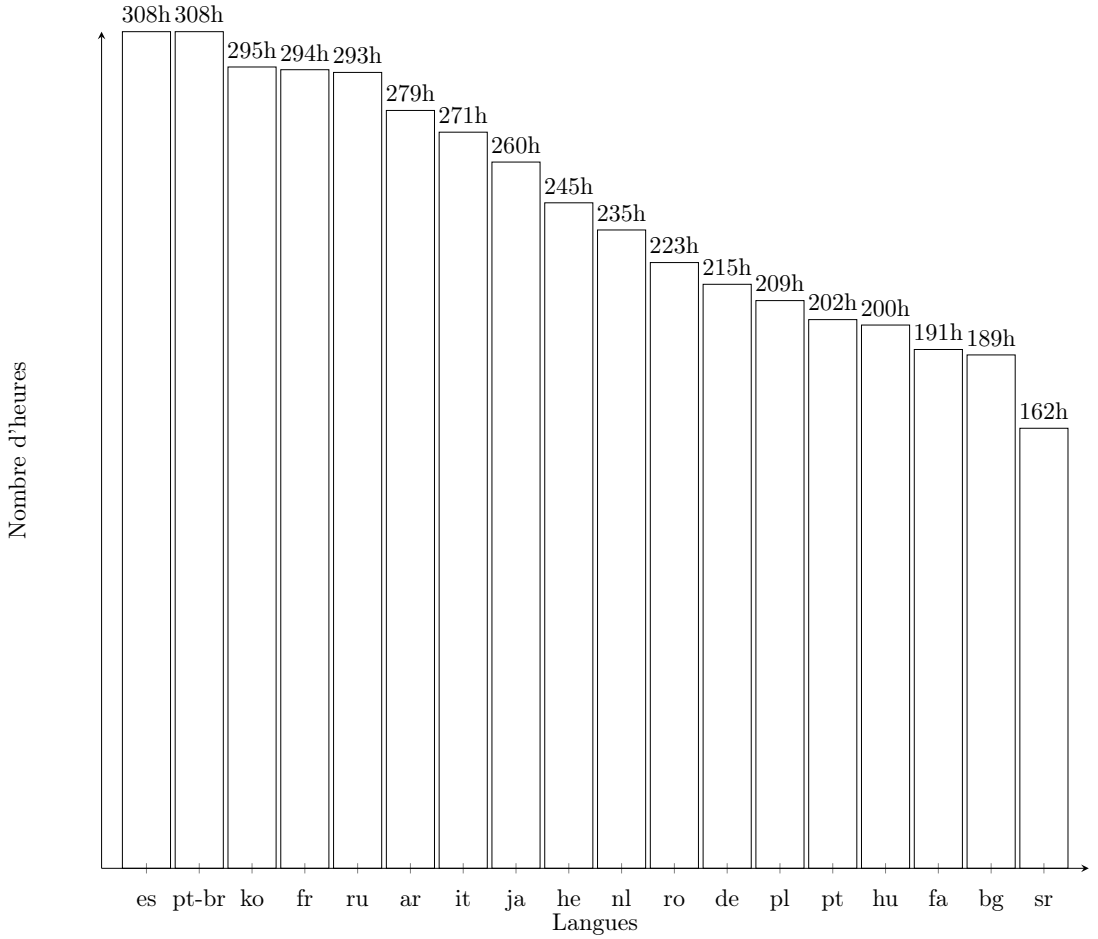


FIGURE 7.6 – Nombre d'heures de vidéos dont la transcription est disponible dans les langues les plus utilisées. Les étiquettes sur l'axe des abscisse représentent les abréviations de langues selon la norme ISO 639-1

transcriptions en arabe vu que notre objectif dans cette partie est de traduire la parole anglaise en arabe. Nous avons collecté 1 119 vidéos en anglais avec des transcriptions en arabe ; ce qui est équivalent à 279 heures.

En analysant le corpus collecté, nous avons trouvé quelques passages où la transcription ne correspond pas bien à ce qui a été prononcé dans la vidéo. Pour y remédier, nous avons proposé une approche classique afin de bien aligner le signal vocal avec la traduction. Cette approche est basée sur les étapes décrites par la figure 7.7.

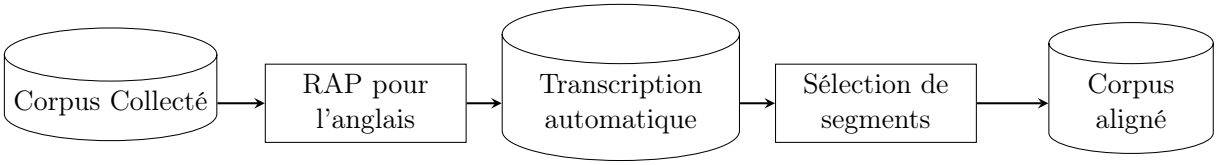


FIGURE 7.7 – Processus d'alignement des données collectées à partir des conférences TED.

— Nous avons lancé une reconnaissance automatique de la parole pour les 1 119 vidéos en

anglais [21] en utilisant un système de reconnaissance pour l'anglais entraîné sur le corpus TED-LIUM 3 [Hernandez *et al.*, 2018]. Les approches utilisées pour entraîner les différents modèles (modèle acoustique et le modèle de langage) dans ce nouveau système sont celles utilisées dans le système de reconnaissance automatique de la parole pour l'anglais du projet AMIS.

— Une fois la transcription en anglais générée, nous avons calculé les taux d'erreur pour chaque segment dans la transcription.

— Enfin, nous avons sélectionné les segments qui ont un WER inférieur strictement à 50%, ce qui réduit le nombre d'heures de 279 heures à 199 heures.
Rappelons que le corpus aligné correspond à une collection de segments en anglais. Afin de récupérer la transcription en arabe de chaque segment, nous avons simplement récupéré la traduction à partir du corpus collecté.

Nous avons divisé le corpus final après le processus d'alignement en trois parties : 195 heures pour l'apprentissage, 2 heures pour la validation et 2 heures pour le test. Dans ce qui suit, nous présentons nos résultats obtenus en traduisant la parole anglaise en texte arabe en utilisant un modèle séquence-à-séquence.

Résultats et discussion

Pour notre modèle séquence-à-séquence, nous avons utilisé la même architecture de modèle que pour la traduction automatique du texte (voir chapitre 6). L'encodeur de ce modèle code les observations acoustiques (des vecteurs MFCC calculés à partir du signal de la parole anglaise) en utilisant deux couches de cellules LSTM bidirectionnelle (bLSTM). Les unités acoustiques (caractères arabes dans notre cas) sont générées en utilisant deux couches de cellules LSTM. Dans le but de modéliser implicitement l'alignement entre les observations acoustiques du signal source et les caractères de la phrase cible, nous avons utilisé le modèle d'attention. L'algorithme d'optimisation *Adaptive moment estimation (Adam)* [Kingma et Ba, 2014] est utilisé pour mettre à jour et estimer les paramètres du modèle avec un *learning rate* de 2×10^{-4}.

La traduction de l'ensemble de validation n'a pas abouti à des résultats acceptables. Deux facteurs peuvent influencer les résultats obtenus :

— La qualité du corpus collecté : le problème majeur dans notre corpus multilingue est d'assurer un bon alignement entre l'audio et la transcription sachant qu'ils sont dans deux langues différentes. Un autre problème pourrait être la qualité de l'audio de ce corpus. Contrairement aux corpus oraux où l'enregistrement est souvent de bonne qualité, l'enregistrement des conférences TED peut contenir des bruits sonores (comme l'applaudissement, le rire, etc.), des hésitations ou des répétitions. Tous ces éléments on un impact négatif sur l'encodage des observations acoustique et sur l'alignement entre ces observations et les unités acoustiques.

— L'architecture du modèle : le problème majeur des modèle séquence-à-séquence est la difficulté d'aligner les longues séquences d'entrée avec les séquences de sortie. Ce constat est particulièrement valable pour le signal audio. Pour un segment d'audio de trois secondes, le nombre d'observations acoustiques extraites à partir de cet audio est estimé à 300. Cela implique que la séquence d'entrée a une taille de 300 vecteurs numériques. Pour remédier à ce problème, les réseaux de neurones convolutif sont utilisés pour réduire la taille de la séquence d'entrée et pour mieux aligner par conséquent, la séquence d'entrée et celle de sortie [Weiss *et al.*, 2017].

21. Ce sont les vidéos en anglais possédant une transcription en arabe

Bien que l'architecture de notre modèle ne soit pas profonde et que nos résultats soient mauvais, ces premiers tests ouvrent une voix de recherche dans le domaine de la traduction automatique de la parole et les facteurs cités ci-dessus représentent un bon point de départ.

7.9 Conclusion et discussion

Nous avons présenté dans ce chapitre nos contributions dans le cadre du projet AMIS. Les deux principaux objectifs du projet étaient de : proposer un système d'aide à la compréhension des vidéos dans des langues étrangères et de proposer une analyse comparative en termes de sentiments entre deux vidéos portant sur le même sujet mais dans deux langues différentes. Atteindre ces deux objectifs implique plusieurs défis scientifiques, à savoir le résumé automatique de vidéos/textes, la reconnaissance automatique de la parole, la traduction automatique et l'analyse de sentiments. Les travaux présentés dans ce chapitre ont porté principalement sur l'interaction entre les systèmes de reconnaissance automatique de la parole et la traduction automatique.

Dans un premier temps, nous avons présenté brièvement le corpus AMIS collecté à partir de contenus en ligne. Ce corpus a été principalement utilisé pour tester le système final et ses différents composants. Le premier défi que nous avons relevé était la récupération des transcriptions de référence pour les données de test du projet AMIS afin de pouvoir évaluer les performances des systèmes de reconnaissance automatique de la parole. Notre approche était d'aligner les vidéos du corpus AMIS avec leurs descriptions récupérées principalement de YouTube et du site d'Euronews. En se basant sur ces transcriptions *approximatives*, nous avons pu calculer une estimation des taux d'erreur produits par les systèmes de reconnaissance automatique de la parole. Les résultats obtenus ont montré des taux d'erreur plus élevés par rapport à ceux calculés sur les corpus de test des systèmes de base. Cela est principalement dû aux entités nommées (nom de personnes, de lieux, etc.) qui ne sont pas reconnues par les systèmes de base. En partant de cette analyse, nous avons proposé une approche pour adapter les systèmes de reconnaissance automatique de la parole aux données collectées dans le cadre du projet AMIS. L'approche était principalement basée sur l'adaptation des vocabulaires de reconnaissance afin de réduire les taux de mots hors vocabulaire. Les résultats obtenus ont montré une réduction absolue de 2,5% dans les taux d'erreur avec les nouveaux systèmes adaptés. Cela a un impact direct sur les performance des systèmes de traduction automatique qui prennent en entrée les transcriptions générées par les systèmes de reconnaissance automatique de la parole afin de traduire les vidéos.

Afin de tester les systèmes de traduction automatique sur le corpus AMIS, une traduction de référence de chaque vidéo est nécessaire. Malheureusement, cette traduction n'était pas disponible ; notre approche pour y remédier a consisté à utiliser une traduction approximative générée par le système de traduction de Google translate. L'évaluation a donné un BLEU acceptable de 26,7%. Ce qui prouve que la traduction finale générée pour chaque vidéo est compréhensible.

L'architecture séquentielle proposée dans le cadre du projet AMIS pour traduire les vidéos souffre du problème de propagation des erreurs entre les différents composants. Ces erreurs peuvent conduire à des textes incompréhensibles vu que le contexte global peut être perdu d'un système à un autre. Pour minimiser ces erreurs entre les composants du système final, nous avons proposé une approche pour traduire directement de la parole avec un seul modèle. L'idée était d'utiliser un modèle séquence-à-séquence qui prend en entrée une séquence d'observations acoustiques dans une langue et génère en sortie une séquence de mots dans une autre langue. Le problème majeur pour entraîner ce modèle est la disponibilité des corpus oraux alignés où le signal acoustique est dans une langue et la transcription dans une autre langue. Nous avons collecté un corpus d'un millier d'heures de parole anglaise et sa transcription dans plusieurs langues (plus de

50 langues). Ce corpus collecté à partir des conférences TED représente une ressource multilingue très utile pour initier la recherche dans le domaine de la traduction automatique de la parole avec les modèles séquence-à-séquence. Bien que notre modèle proposé n'ait pas donné de bon résultats, nous avons présenté dans le cadre de cette thèse les éventuelles problèmes qui n'ont pas été résolus et qui peuvent constituer un bon point de départ pour améliorer les performances des systèmes de traduction de la parole.

Conclusion et perspectives

Le travail réalisé dans le cadre de cette thèse est inscrit dans le cadre du projet AMIS *Access to Multilingual Information and opinionS*. AMIS est un projet européen dont l'objectif principal est d'aider les personnes à comprendre l'idée générale d'une vidéo dans une langue étrangère. Notre travail s'est focalisé sur la traduction automatique de la parole arabe. Nous avons commencé par le développement du système ALASR : un système de reconnaissance automatique de la parole pour la langue arabe standard. Ce même système a été adapté pour reconnaître une variante du dialecte algérien : le dialecte algérois connu pour être le dialecte de la région du centre de l'Algérie. Ensuite, dans le but de traduire le texte généré par le système de reconnaissance automatique de la parole, nous avons réalisé une étude comparative entre l'approche statistique à base de segments et l'approche neuronale. Enfin, nous avons travaillé sur l'adaptation des systèmes de traduction automatique pour traiter le *code-switching*, un phénomène très répandu dans le monde arabe.

Le système ALASR, *Arabic Loria Automatic Speech Recognition system*, est un système de reconnaissance automatique de la parole arabe à base de trois composants : un modèle acoustique, un modèle de langage et un lexique. Notre modèle acoustique est un modèle DNN-HMM où pour chaque observation acoustique O, on essaie de prédire l'état du modèle de Markov caché qui lui correspond au mieux. L'ensemble de trois états du modèle de Markov caché modélise un triphone (un phonème dans son contexte gauche et droit). Nous avons testé principalement deux architectures neuronales : une basée sur les modèles perceptron multicouche, *multilayer perceptron (MLP)* [Hinton *et al.*, 2012a] et l'autre est basée sur les modèles à retard temporel, *Time Delay Neural Network (TDNN)* [Peddinti *et al.*, 2015]. La différence entre les deux architectures est la modélisation du contexte de chaque observation acoustique pour prédire les unités acoustiques. En effet, dans le modèle à base de MLP, la prédiction des états du modèle de Markov caché est basée sur une fenêtre de contexte fixe (la concaténation de plusieurs observations acoustiques) tandis que dans le modèle TDNN, on utilise plusieurs fenêtres de contexte tout en commençant par un contexte étroit au niveau de la couche d'entrée et des contextes plus larges sont utilisés dans les couches supérieures. Cela permet aux couches supérieures d'apprendre des relations temporelles entre les observations acoustiques plus larges. L'apprentissage des modèles DNN-HMM nécessite une grande quantité de données étiquetées de la forme (observation acoustique, unité acoustique) qui est souvent non disponible. L'alignement forcé est l'approche la plus utilisée pour générer ces données. Il consiste à entraîner un modèle de base et à l'utiliser pour aligner les données orales avec les unités acoustiques. Nous avons entraîné dans notre cas plusieurs modèles GMM-HMM en changeant les critères d'apprentissage et en appliquant une série de transformations afin de trouver le meilleur modèle et de l'utiliser pour initialiser l'apprentissage des modèles DNN-HMM.

Au niveau de la modélisation du langage, nous avons comparé plusieurs modèles n-grammes où nous avons varié l'ordre n du modèle, testé plusieurs techniques d'élagage et proposé une phase de normalisation de données vu que l'orthographe de l'arabe est déroutante du fait que

l'écriture de certains mots peut être simplifiée. En ce qui concerne le lexique de prononciations, le problème majeur auquel nous étions confronté, était la restitution des voyelles dans le texte afin de générer les variantes de prononciation des mots. Bien qu'une simple décomposition des mots en lettres soit possible en arabe pour générer la prononciation d'un mot, nous avons montré que cette solution n'assure pas une bonne modélisation acoustique pour les voyelles. Pour cela, nous avons proposé plusieurs approches pour la modélisation explicite des voyelles courtes, longues et la gémination. Le système final nous a permis d'obtenir un taux d'erreur de 12,7% sur un ensemble de test contenant 5 heures de paroles arabes. Ce système est celui obtenu avec un modèle acoustique basé sur le réseau de neurones TDNN et entraîné avec un corpus composé de 52 heures de parole, un modèle de langage 2-gramme sans élagage entraîné sur 1 milliard de mots et un lexique de prononciations où nous avons modélisé explicitement les voyelles courtes, longues et la gémination dans le texte. Nous avons également inclus une phase de *rescoring* qui consistait à réévaluer la sortie du système avec un modèle de langage 4-gramme, et une phase de normalisation de l'écriture de *hamza* ء au-dessus/au-dessous de la lettre *Alif* ا. Cela nous a permis d'obtenir une amélioration absolue par rapport au score initial de 1,2% sur l'ensemble de validation et de 0,7% sur l'ensemble de test.

Étant donné que l'arabe standard n'est pas la langue maternelle dans le monde arabe et qu'on utilise plutôt dans les conversations quotidiennes une autre variante de langue connue sous le nom de dialecte, nous avons travaillé sur l'adaptation du système ALASR pour reconnaître une variante du dialecte algérien : le dialecte algérois. Du fait que le locuteur algérien exploite toutes ces ressources langagières disponibles dans son paysage linguistique pour s'exprimer, deux principaux phénomènes caractérisent son dialecte : l'emprunt et le *code-switching*.
Les mots empruntés d'autres langues, principalement du français, tiennent une place importante dans le vocabulaire du dialecte algérien. Le *code-switching* est défini comme l'utilisation de plus d'une langue par un locuteur dans un même énoncé ou discours. Le problème majeur pour le traitement automatique du dialecte est le manque de ressources vocales et textuelles. Notre approche pour y remédier était de tirer profit des données d'autres langues qui influencent ce dialecte, à savoir l'arabe standard et le français. Avec l'émergence des réseaux sociaux, ces derniers représentent une source primordiale pour récupérer des ressources textuelles pour le dialecte algérien. Notre modèle de langage final était l'interpolation linéaire entre plusieurs modèles de langage, à savoir le modèle utilisé dans le système ALASR et d'autres modèles entraînés sur une collection de commentaires provenant de YouTube [Abidi et Smaïli, 2017, Abidi *et al.*, 2017] et de conversations en dialecte [Meftouh *et al.*, 2015, Meftouh *et al.*, 2018]. En ce qui concerne les données orales, nous avons construit le corpus ADIC : *Algerian DIalect Corpus*. C'est un corpus contenant 6 heures de parler algérien avec leur transcription ; nous l'avons utilisé pour entraîner un modèle acoustique de base et pour tester les performances du système ALASR après son adaptation au dialecte. Cependant, l'apprentissage des modèles DNN-HMM nécessite une grande quantité de données acoustiques, quantité qui n'est pas disponible pour le dialecte. En partant du constat que l'usage des mots empruntés dans le dialecte algérien implique l'utilisation de nouveaux sons en plus des sons de l'arabe standard et qui sont souvent utilisés en français, nous avons enrichi le corpus ADIC avec des données orales d'arabe et de français. Nous avons proposé différentes approches pour intégrer ces données dans le processus de l'apprentissage du modèle acoustique du dialecte algérien :

— L'apprentissage **multilingue** qui consiste à apprendre un seul modèle acoustique à partir d'un mélange de données de plusieurs langues : le dialecte algérien, l'arabe standard et le français. Techniquement parlant, toutes les couches de notre modèle neuronal y compris la couche de sortie qui estime les probabilités des unités acoustiques de chaque langue

sont partagées entre les trois langues.

— L'apprentissage **multitâche** qui consiste à apprendre un seul modèle neuronal pour reconnaître les trois langues à la fois. Dans ce cas, toutes les couches cachées du modèle sont partagées entre les trois langues et les unités acoustiques de chaque langue sont modélisées séparément ; il n'y a aucun partage au niveau de la couche de sortie.

— L'apprentissage par **transfert de connaissance** qui consiste à entraîner un modèle initial sur une tâche (la reconnaissance de l'arabe et du français) avec une grande quantité de données et adapter le modèle obtenu pour une nouvelle tâche (la reconnaissance du dialecte) où peu de données d'apprentissage sont disponibles. Concrètement, le processus d'adaptation consiste à supprimer la dernière couche de sortie et à garder seulement les n premières couches cachées du modèle initial. Une nouvelle couche spécifique au dialecte algérien est ajoutée au-dessus des couches cachées du modèle initial.

Nous avons trouvé que le meilleur moyen pour exploiter les données d'autres langues dans la modélisation acoustique du dialecte est d'apprendre le réseau de neurones sur un mélange de données de plusieurs langues tout en partageant toutes les couches du réseau de neurones entre ces différentes langues (l'apprentissage multilingue). Nous avons obtenu une amélioration absolue de 6,7% par rapport au modèle de base où seulement les données dialectales ont été utilisées pour l'apprentissage du modèle acoustique et du modèle de langage.

Afin de traduire le texte généré par le système ALASR, nous avons travaillé sur la traduction automatique de l'arabe vers l'anglais tout en considérant les aspects relatifs au dialecte algérien et en utilisant un corpus parallèle arabe-anglais. L'utilisation d'un tel corpus est justifié par le fait que les corpus parallèles dialectaux sont peu disponibles voire non disponibles. Nous avons mené une étude comparative entre l'approche statistique à base de segments et l'approche neuronale dans un cadre où peu de données d'apprentissage sont disponibles. Nous avons également travaillé sur la traduction des documents contenant du *code-switching* entre l'arabe et l'anglais.

Les résultats de l'étude comparative entre les deux approches les plus dominantes dans le domaine de la traduction automatique ont montré que dans le cadre où peu de données parallèles sont disponibles, l'approche neuronale a besoin de composants externes et de mettre en place une architecture plus aboutie (utilisation du modèle d'attention, ajustement des probabilités pour traiter les mots inconnus/rares, etc.) pour se rapprocher des performances de l'approche statistique.

Le *code-switching* est un phénomène très répandu dans les dialectes arabes. Nous avons montré que son existence n'est pas restreinte aux communications informelles, il est bien présent dans les procédures des institutions multilingues. En se basant sur les documents officiels des nations unies, nous avons construit une ressource parallèle consistant en un texte source *code-switché* et sa traduction en arabe standard pur et en anglais pur. Cette dernière a été évaluée manuellement afin de s'assurer de sa qualité. Nous avons également travaillé sur l'adaptation des systèmes de traduction automatique pour traiter le *code-switching* en proposant différentes stratégies de traduction. Nous avons trouvé pour l'approche neuronale qu'en recopiant directement les segments en anglais dans la sortie du système, on peut avoir de bien meilleurs résultats. Pour l'approche statistique, nous avons trouvé que l'apprentissage du modèle de traduction sur un corpus *code-switché* artificiel a donné de meilleurs résultats de traduction du texte *code-switché*.

Dans le cadre du projet AMIS, nous avons décidé d'utiliser une architecture séquentielle afin de traduire des vidéos arabes en anglais. Dans ce cadre, nous avons utilisé la sortie du système de reconnaissance automatique de la parole comme entrée du système de traduction. L'évaluation de ce modèle a été effectuée sur le corpus AMIS collecté à partir de contenus en ligne. Le premier défi que nous avons relevé était la récupération des transcriptions de référence pour le corpus AMIS afin de pouvoir évaluer les performances du système ALASR. Nous avons proposé

une approche pour aligner automatiquement les vidéos du corpus AMIS avec leurs descriptions récupérées principalement de YouTube et du site d'Euronews. En se basant sur ces transcriptions *approximatives*, nous avons estimé les taux d'erreur produits par le système de reconnaissance automatique de la parole. Les résultats obtenus ont montré des taux d'erreur plus élevés par rapport à ceux calculés sur le corpus de test du système ALASR. Cela a un impact direct sur la traduction automatique dans le modèle séquentiel puisque ces erreurs peuvent conduire à des textes entachés d'erreurs où le contexte global peut être perdu d'un système à un autre.

Dans le but de minimiser les taux d'erreur sur le corpus AMIS, nous avons proposé une approche pour adapter le système de reconnaissance automatique de la parole aux données collectées dans le cadre du projet AMIS. L'approche était principalement basée sur l'adaptation des vocabulaires de reconnaissance afin de réduire les taux de mots hors vocabulaire. En effet, nous avons observé qu'une grande partie des erreurs est due aux entités nommées (nom des personnes, des lieux, etc.) qui n'existent pas dans le vocabulaire et qui n'ont pas été reconnues par conséquent par le système ALASR. Les résultats obtenus après l'adaptation du vocabulaire ont montré une réduction absolue de 2,5% dans le taux d'erreur.
En traduisant la sortie du système de reconnaissance automatique de la parole, nous avons obtenu un BLEU acceptable de 26,7%. Ce qui prouve que la traduction finale générée pour chaque vidéo du corpus AMIS est compréhensible.

Bien que notre approche pour adapter le vocabulaire du système de reconnaissance automatique de la parole ait amélioré la sortie de ce dernier, elle ne règle pas le problème de propagation des erreurs dans le modèle séquentiel. En partant de ce constat, nous avons proposé d'utiliser des systèmes de bout en bout pour traduire directement la parole. Ces modèles sont souvent utilisés dans la littérature dans le cadre de la reconnaissance automatique de la parole ou de la traduction automatique du texte. Notre proposition était de tirer profit de l'application de ces modèles dans les deux domaines afin de développer un système prenant en entrée un signal de la parole dans une langue et générer en sortie une séquence de mots dans une autre langue. L'apprentissage de ce modèle nécessite une grande quantité de données parallèles de la forme (observations acoustiques langue A, séquence de mots langue B) qui n'est pas disponible. Pour y remédier, nous avons construit une ressource multilingue en collectant des milliers d'heures de parole anglaise et leurs transcriptions dans plusieurs langues (plus de 50 langues) à partir des conférences TED.

En utilisant cette ressource, nous avons proposé un modèle séquence-à-séquence qui prend en entrée une séquence d'observations acoustiques de la parole anglaise, il la projette dans un espace multidimensionnel en utilisant deux couches de cellules bLSTM et il génère en sortie une séquence de caractères arabes correspondant au signal d'entrée. Ce modèle n'a pas abouti à des résultats de traduction comparables à ceux obtenus avec le modèle séquentiel. Deux pistes de recherche peuvent être exploitées afin de booster le modèle de bout en bout :

— Améliorer la qualité du corpus collecté : la voix dans les conférences TED n'est pas forcément de bonne qualité. Il existe des passages qui ne sont pas pertinents pour la traduction automatique de la parole comme l'hésitation, la répétition et les bruits sonores (l'applaudissement, le rire, etc.). Ces passages ont un impact direct sur l'alignement entre le signal acoustique et la transcription qui sont dans notre cas dans deux langues différentes. L'approche que nous avons proposée pour aligner le signal acoustique avec la transcription n'assure pas que la traduction correspond bien à ce qui a été prononcé dans l'audio. En effet, pour aligner le signal acoustique avec sa traduction, nous avons sélectionné les segments en anglais qui correspondent au mieux à ce qui a été prononcé ; et vu la nature parallèle de notre ressource, nous avons sélectionné la traduction en arabe de ces segments en anglais. Dans ce cas et pour aligner un signal acoustique avec sa traduction,

nous nous sommes seulement focalisés sur la transcription en anglais, nous n'avons pas pris en considération sa traduction. Il serait intéressant de s'assurer que les segments en anglais et leur traduction sont bien alignés via un processus d'alignement forcé.

— Proposer une architecture plus profonde/complexe : le modèle que nous avons proposé est typiquement utilisé pour la traduction automatique du texte où les séquences d'entrée sont de taille réduite (des séquences de 50 mots en moyenne). Dans le cas de la parole, la taille de cette séquence d'entrée est beaucoup plus grande ; pour un segment d'audio de trois secondes, le nombre d'observations acoustiques est estimé à 300. Cela a un impact direct sur le modèle d'attention qui est souvent utilisé pour aligner les séquences d'entrée avec les séquences de sortie. Plus la taille de la séquence d'entrée est grande, plus le modèle d'attention aura du mal à bien aligner cette séquence avec celle de sortie. Une solution à ce problème serait d'utiliser des encodeurs basés sur la combinaison des réseaux de neurones convolutifs avec des réseaux de neurones récurrents. L'objectif des réseaux de neurones convolutifs est de réduire la taille de la séquence d'observations acoustiques ce qui permet au modèle d'attention de bien capturer les liens entre la séquence d'entrée et celle de sortie.

Toutes les ressources développées dans le cadre de cette thèse : le corpus ADIC, le corpus parallèle *code-switché* et le corpus multilingue TED, sont accessibles à partir du site officiel de l'équipe SMarT.

Annexe A

Nous présentons ici la liste des abréviations des mots utilisées dans le processus de normalisation de données textuelles arabes ainsi que leur signification.

Abréviation	**Remplacée par**	**Traduction**
%	في المائة	Pourcent
ص	صفحة	Page
ت غ	توقيت غرينتش	GMT
ت	تَاريخ	Date
هـ	هجري	*Hijri* (calendrier islamique)
م	ميلَادي	Calendrier grégorien
د.	دكتور	Docteur
أ . د	الَاُستَاذ الدكتور	Professeur
د . ك	دولَار كندي	Dollar canadien
س	سَاعة	Heure
د	دقيقة	Minute
ث	ثَانية	Second
كم ـ كلم	كيلومتر	Kilomètre
م	متر	Mètre
سم	سنتمتر	Centimètre
مّ	ميليمتر	Millimètre
م^2	متر مربع	Mètre carré
ق م	قبل الميلَاد	B.C.
الخ	اِلَى آخره	Etc.

(a) Liste des abréviations avec leur signification.

Erreur d'orthographe	**Remplacée par**	**Traduction**
ف المائة	في المائة	Pourcent
بالم إة	بالمائة	Pourcent
بالما إة	بالمائة	Pourcent
الثلثاء	الثلاثاء	Mardi
فى	في	Dans
هى	هي	Elle
الى	إلى	À/Vers
حتي	حتى	Jusqu'à
او	أو	Ou
اي	أي	N'importe

(b) Liste des erreurs d'orthographe les plus fréquentes.

Table A.1 – Liste des abréviations et des erreurs d'orthographe les plus fréquentes et leur correction.

Bibliographie

[Abdelali *et al.*, 2016] Abdelali, A., Darwish, K., Durrani, N. et Mubarak, H. (2016). Farasa : A fast and furious segmenter for arabic. *In HLT-NAACL Demos*.

[Abidi, 2019] Abidi, K. (2019). *Automatic building of multilingual resources from social networks : application to Maghrebi dialects*. Theses, Université de Lorraine.

[Abidi *et al.*, 2017] Abidi, K., Menacer, M. A. et Smaili, K. (2017). CALYOU : A Comparable Spoken Algerian Corpus Harvested from YouTube. *In 18th Annual Conference of the International Communication Association (Interspeech)*, Conference of the International Communication Association (Interspeech), Stockholm, Sweden.

[Abidi et Smaïli, 2017] Abidi, K. et Smaïli, K. (2017). An empirical study of the Algerian dialect of Social network. *In ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*, Casablanca, Morocco.

[Abidi et Smaïli, 2018] Abidi, K. et Smaïli, K. (2018). An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings. *In 11th edition of the Language Resources and Evaluation Conference, LREC 2018*.

[Abo-Bakr *et al.*, 2008] Abo-Bakr, H., Shaalan, K. et Ziedan, I. (2008). A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. *In The 6th International Conference on Informatics and Systems (INFOS2008)*, Cairo, Egypt. Faculty of Comptuers and Information, Faculty of Comptuers and Information.

[Afify *et al.*, 2006] Afify, M., Sarikaya, R., kwang Jeff Kuo, H., Besacier, L. et Gao, Y. (2006). On the use of morphological analysis for dialectal arabic speech recognition. *In Proceedings of ICSLP'06*, pages 277–280.

[Ali *et al.*, 2014] Ali, A., Mubarak, H. et Vogel, S. (2014). Advances in dialectal arabic speech recognition : A study using twitter to improve egyptian ASR. *In International Workshop on Spoken Language Translation (IWSLT 2014)*.

[Ali *et al.*, 2019] Ali, A., Shon, S., Samih, Y., Mubarak, H., Abdelali, A., Glass, J., Renals, S. et Choukri, K. (2019). The mgb-5 challenge : Recognition and dialect identification of dialectal arabic speech.

[Ali *et al.*, 2014] Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S. et Glass, J. (2014). A complete kaldi recipe for building arabic speech recognition systems. *In 2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 525–529.

[Alkanhal *et al.*, 2012] Alkanhal, M. I., Al-Badrashiny, M. A., Alghamdi, M. M. et Al-Qabbany, A. O. (2012). Automatic stochastic arabic spelling correction with emphasis on space insertions and deletions. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):2111–2122.

[Alqudsi *et al.*, 2012] Alqudsi, A., Omar, N. et Shaker, K. (2012). Arabic machine translation : A survey. *Artificial Intelligence Review*, 42.

[AlShenaifi *et al.*, 2015] AlShenaifi, N., AlNefie, R., Al-Yahya, M. et Al-Khalifa, H. (2015). Arib@ qalb-2015 shared task : A hybrid cascade model for arabic spelling error detection and correction. *In ANLP Workshop 2015*, page 127.

[Amari, 1998] Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

[Amodei *et al.*, 2016] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. *et al.* (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. *In International conference on machine learning*, pages 173–182.

[Anastasakos *et al.*, 1996] Anastasakos, T., McDonough, J., Schwartz, R. et Makhoul, J. (1996). A compact model for speaker-adaptive training. *In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 1137–1140 vol.2.

[Arthur *et al.*, 2016] Arthur, P., Neubig, G. et Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. *CoRR*, abs/1606.02006.

[Attia *et al.*, 2014] Attia, M., Al-Badrashiny, M. et Diab, M. (2014). Gwu-hasp : Hybrid arabic spelling and punctuation corrector. *In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 148–154.

[Attia *et al.*, 2012] Attia, M., Pecina, P., Samih, Y., Shaalan, K. et van Genabith, J. (2012). Improved spelling error detection and correction for arabic. *In The International Conference on Computational Linguistics (COLING)*, pages 103–112, Mumbai, India.

[Auer, 1999] Auer, P. (1999). From codeswitching via language mixing to fused lects : Toward a dynamic typology of bilingual speech. *International journal of bilingualism*, 3(4):309–332.

[Bahdanau *et al.*, 2014] Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

[Bahi et Sellami, 2003] Bahi, H. et Sellami, M. (2003). A hybrid approach for arabic speech recognition. page 107.

[Bahl *et al.*, 1986] Bahl, L., Brown, P., de Souza, P. et Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. *In ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52.

[Banerjee et Lavie, 2005] Banerjee, S. et Lavie, A. (2005). Meteor : An automatic metric for mt evaluation with improved correlation with human judgments. *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

[Baniata *et al.*, 2018] Baniata, L. H., Park, S. et Park, S.-B. (2018). A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

[Baum et Eagon, 1967] Baum, L. E. et Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363.

[Bengio *et al.*, 2003] Bengio, Y., Ducharme, R., Vincent, P. et Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

[Bentivogli *et al.*, 2016] Bentivogli, L., Bisazza, A., Cettolo, M. et Federico, M. (2016). Neural versus phrase-based machine translation quality : a case study. *CoRR*, abs/1608.04631.

[Bernath *et al.*, 2018] Bernath, C., Alvarez, A., Arzelus, H. et Martínez, C. D. (2018). Exploring e2e speech recognition systems for new languages. *In IberSPEECH*, pages 102–106.

[Boisvert, 2006] Boisvert, M. (2006). Techniques for vocabulary selection and word weighting in language models.

[Bottou, 1998] Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142.

[Bourouba *et al.*, 2006] Bourouba, H., Djemili, R., Bedda, M. et Snani, C. (2006). New hybrid system (supervised classifier/hmm) for isolated arabic speech recognition. *In 2006 2nd International Conference on Information Communication Technologies*, volume 1, pages 1264–1269.

[Brown *et al.*, 1990] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. et Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

[Brown *et al.*, 1993] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. et Mercer, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19(2):263–311.

[Bullock *et al.*, 2014] Bullock, B. E., Hinrichs, L. et Toribio, A. J. (2014). World englishes, code-switching, and convergence. *Oxford Handbook of World Englishes*.

[Callison-Burch *et al.*, 2006] Callison-Burch, C., Osborne, M. et Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. *In 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

[Carpuat, 2014] Carpuat, M. (2014). Mixed language and code-switching in the canadian hansard. *In Proceedings of the first workshop on computational approaches to code switching*, pages 107–115.

[Cavnar *et al.*, 1994] Cavnar, W. B., Trenkle, J. M. *et al.* (1994). N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175.

[Ceausu et Tufis, 2012] Ceausu, A. et Tufis, D. (2012). Addressing smt data sparseness when translating into morphologically-rich languages.

[Chan *et al.*, 2015] Chan, W., Jaitly, N., Le, Q. V. et Vinyals, O. (2015). Listen, attend and spell. *CoRR*, abs/1508.01211.

[Chen et Goodman, 1996] Chen, S. F. et Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *In Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

[Cho *et al.*, 2014a] Cho, K., van Merrienboer, B., Bahdanau, D. et Bengio, Y. (2014a). On the properties of neural machine translation : Encoder-decoder approaches. *CoRR*, abs/1409.1259.

[Cho *et al.*, 2014b] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H. et Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

[Choueiter *et al.*, 2006] Choueiter, G., Povey, D., Chen, S. F. et Zweig, G. (2006). Morpheme-based language modeling for arabic lvcsr. *In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

[Choukri *et al.*, 2004] Choukri, K., Nikkhou, M. et Paulsson, N. (2004). Network of data centres (NetDC) : BNSC - an Arabic broadcast news speech corpus. *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

[Coughlin, 2003] Coughlin, D. A. (2003). Correlating automated and human assessments of machine translation quality.

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

[Davis et Mermelstein, 1980] Davis, S. et Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

[Diehl *et al.*, 2012] Diehl, F., Gales, M., Tomalin, M. et Woodland, P. (2012). Morphological decomposition in arabic asr systems. *Computer Speech & Language*, 26(4):229 – 243.

[Diehl *et al.*, 2009] Diehl, F., Gales, M. J., Tomalin, M. et Woodland, P. C. (2009). Morphological analysis and decomposition for arabic speech-to-text systems. *In Interspeech*, pages 2675–2678.

[Douib *et al.*, 2016] Douib, A., Langlois, D. et Smaili, K. (2016). Genetic-based decoder for statistical machine translation.

[Duchi *et al.*, 2011] Duchi, J., Hazan, E. et Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159.

[Dugast *et al.*, 2007] Dugast, L., Senellart, J. et Koehn, P. (2007). Statistical post-editing on systran's rule-based translation system. *In Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 220–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Eisele et Chen, 2010] Eisele, A. et Chen, Y. (2010). Multiun : A multilingual corpus from united nation documents. *In* Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K. et Chair), N. C. C., éditeurs : *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA).

[El-Desoky *et al.*, 2009] El-Desoky, A., Gollan, C., Rybach, D., Schlüter, R. et Ney, H. (2009). Investigating the use of morphological decomposition and diacritization for improving arabic lvcsr. *In Interspeech*, pages 2679–2682.

[Elmahdy *et al.*, 2012] Elmahdy, M., Hasegawa-Johnson, M. et Mustafawi, E. (2012). A baseline speech recognition system for levantine colloquial arabic. *Proceedings of ESOLEC*.

[Elmahdy *et al.*, 2014] Elmahdy, M., Hasegawa-Johnson, M. et Mustafawi, E. (2014). Development of a tv broadcasts speech recognition system for qatari arabic. *In LREC*, pages 3057–3061.

[Fohr *et al.*, 2017] Fohr, D., Mella, O. et Illina, I. (2017). New Paradigm in Speech Recognition : Deep Neural Networks. *In IEEE International Conference on Information Systems and Economic Intelligence*, Marrakech, Morocco.

[Forney, 1973] Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

[Gales, 1998] Gales, M. J. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98.

[Galliano *et al.*, 2009] Galliano, S., Gravier, G. et Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. *In Proceddings of Interspeech, Brighton (United Kingdom)*.

[Garg *et al.*, 2017] Garg, S., Parekh, T. et Jyothi, P. (2017). Dual language models for code mixed speech recognition. *CoRR*, abs/1711.01048.

[Gibson et Hain, 2006] Gibson, M. et Hain, T. (2006). Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. *In Ninth international conference on spoken language processing*.

[Gopinath, 1998] Gopinath, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. *In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 2, pages 661–664 vol.2.

[Graves *et al.*, 2006] Graves, A., Fernández, S., Gomez, F. et Schmidhuber, J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. *In Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

[Graves et Jaitly, 2014] Graves, A. et Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. *In International conference on machine learning*, pages 1764–1772.

[Guella, 2011] Guella, N. (2011). Emprunts lexicaux dans des dialectes arabes algériens. *Synergies Monde Arabe*, 8:81–88.

[Habash, 2007] Habash, N. (2007). Syntactic preprocessing for statistical machine translation.

[Haeb-Umbach et Ney, 1992] Haeb-Umbach, R. et Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. *In [Proceedings] ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 13–16 vol.1.

[Hannun *et al.*, 2014] Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. et Ng, A. Y. (2014). Deep speech : Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

[Harrat *et al.*, 2014] Harrat, S., Meftouh, K., Abbas, M. et Smaïli, K. (2014). Grapheme to phoneme conversion - an Arabic dialect case. *In Spoken Language Technologies for Under-resourced Languages*.

[Hassan *et al.*, 2014] Hassan, Y., Aly, M. et Atiya, A. (2014). Arabic spelling correction using supervised learning. *arXiv preprint arXiv :1409.8309*.

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

[Hernandez *et al.*, 2018] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N. A. et Estève, Y. (2018). TED-LIUM 3 : twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699.

[Hinton *et al.*, 2012a] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B. *et al.* (2012a). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

[Hinton *et al.*, 2012b] Hinton, G., Srivastava, N. et Swersky, K. (2012b). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

[Hutchins, 2004a] Hutchins, W. J. (2004a). The georgetown-IBM experiment demonstrated in january 1954. *In Machine Translation : From Real Users to Research*, pages 102–114. Springer.

[Hutchins, 2004b] Hutchins, W. J. (2004b). The georgetown-ibm experiment demonstrated in january 1954. *In Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer.

[Imai *et al.*, 1995] Imai, T., Ando, A. et Miyasaka, E. (1995). A new method for automatic generation of speaker-dependent phonological rules. *In 1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 864–867 vol.1.

[Isabelle *et al.*, 2017] Isabelle, P., Cherry, C. et Foster, G. F. (2017). A challenge set approach to evaluating machine translation. *CoRR*, abs/1704.07431.

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

[Joshi, 1982] Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. *In Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

[Jouvet *et al.*, 2017] Jouvet, D., Langlois, D., Menacer, M. A., Fohr, D., Mella, O. et Smaïli, K. (2017). About vocabulary adaptation for automatic speech recognition of video data. *In ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing*, pages 1–5, Casablanca, Morocco.

[Jouvet *et al.*, 2018] Jouvet, D., Langlois, D., Menacer, M. A., Fohr, D., Mella, O. et Smaïli, K. (2018). Adaptation of speech recognition vocabularies for improved transcription of YouTube videos. *Journal of International Science and General Applications*, 1(1):1–9.

[Kalchbrenner et Blunsom, 2013] Kalchbrenner, N. et Blunsom, P. (2013). Recurrent continuous translation models. Seattle. Association for Computational Linguistics.

[Khasawneh *et al.*, 2004] Khasawneh, M., Assaleh, K., Sweidan, W. et Haddad, M. (2004). The application of polynomial discriminant function classifiers to isolated arabic speech recognition. *In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 3077–3081 vol.4.

[Khurana et Ali, 2016] Khurana, S. et Ali, A. (2016). Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition : Mgb-2 challenge. pages 292–298.

[Kingma et Ba, 2014] Kingma, D. P. et Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

[Kirchhoff *et al.*, 2002] Kirchhoff, K. *et al.* (2002). Novel speech recognition models for arabic. *In Johns-Hopkins University summer research workshop*.

[Knight, 1999] Knight, K. (1999). Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615.

[Knight et Koehn, 2003] Knight, K. et Koehn, P. (2003). What's new in statistical machine translation. *In Companion Volume of the Proceedings of HLT-NAACL 2003 - Tutorial Abstracts*, pages 5–5.

[Koehn, 2004] Koehn, P. (2004). Pharaoh : A beam search decoder for phrase-based statistical machine translation models. *In* Frederking, R. E. et Taylor, K. B., éditeurs : *Machine*

Translation : From Real Users to Research, pages 115–124, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Koehn, 2009] Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

[Koehn *et al.*, 2007a] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. *et al.* (2007a). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

[Koehn *et al.*, 2007b] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007b). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

[Koehn et Knowles, 2017] Koehn, P. et Knowles, R. (2017). Six challenges for neural machine translation. *CoRR*, abs/1706.03872.

[Koehn *et al.*, 2003] Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical Phrase-based Translation. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

[Koźbiał et Leszczuk, 2019] Koźbiał, A. et Leszczuk, M. (2019). Collection, analysis and summarization of video content. *In* Choroś, K., Kopel, M., Kukla, E. et Siemiński, A., éditeurs : *Multimedia and Network Information Systems*, pages 405–414, Cham. Springer International Publishing.

[Lample *et al.*, 2017] Lample, G., Denoyer, L. et Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

[Landis et Koch, 1977] Landis, J. R. et Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

[Langlais *et al.*, 2007] Langlais, P., Patry, A. et Gotti, F. (2007). A greedy decoder for phrase-based statistical machine translation.

[Laurent *et al.*, 2009] Laurent, A., Deléglise, P. et Meignier, S. (2009). Grapheme to phoneme conversion using an smt system. *In Tenth Annual Conference of the International Speech Communication Association*.

[Lee *et al.*, 2003] Lee, Y.-S., Papineni, K., Roukos, S., Emam, O. et Hassan, H. (2003). Language model based arabic word segmentation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 399–406. Association for Computational Linguistics.

[Lui et Baldwin, 2012] Lui, M. et Baldwin, T. (2012). langid. py : An off-the-shelf language identification tool. *In Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

[Luong *et al.*, 2014] Luong, T., Sutskever, I., Le, Q. V., Vinyals, O. et Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *CoRR*, abs/1410.8206.

[Manning et Schütze, 1999] Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[May, 2014] May, J. (2014). An arabizi-english social media statistical machine translation system.

[Meftouh *et al.*, 2015] Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M. et Smaili, K. (2015). Machine translation experiments on PADIC : A parallel Arabic DIalect corpus. *In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

[Meftouh *et al.*, 2018] Meftouh, K., Harrat, S. et Smaïli, K. (2018). Padic : extension and new experiments.

[Menacer *et al.*, 2019] Menacer, M., Langlois, D., Jouvet, D., Fohr, D., Mella, O. et Smaïli, K. (2019). Machine Translation on a parallel Code-Switched Corpus. *In Canadian AI 2019 - 32nd Conference on Canadian Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Ontario, Canada.

[Menacer *et al.*, 2017a] Menacer, M. A., Langlois, D., Mella, O., Fohr, D., Jouvet, D. et Smaïli, K. (2017a). Is statistical machine translation approach dead ? *In ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*, pages 1–5, Casablanca, Morocco. ISGA.

[Menacer *et al.*, 2017b] Menacer, M. A., Mella, O., Fohr, D., Jouvet, D., Langlois, D. et Smaïli, K. (2017b). An enhanced automatic speech recognition system for Arabic. *In The third Arabic Natural Language Processing Workshop - EACL 2017*, Arabic Natural Language Processing Workshop - EACL 2017, Valencia, Spain.

[Menacer *et al.*, 2017c] Menacer, M. A., Mella, O., Fohr, D., Jouvet, D., Langlois, D. et Smaïli, K. (2017c). Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. *In ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics*, pages 1–8, Dubai, United Arab Emirates.

[Mikolov *et al.*, 2010] Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J. et Khudanpur, S. (2010). Recurrent neural network based language model. *In Eleventh annual conference of the international speech communication association.*

[Mikolov *et al.*, 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. et Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv :1310.4546.*

[Mohamed *et al.*, 2012] Mohamed, E., Mohit, B. et Oflazer, K. (2012). Transforming standard arabic to colloquial arabic. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2*, pages 176–180. Association for Computational Linguistics.

[Mohri *et al.*, 2008] Mohri, M., Pereira, F. et Riley, M. (2008). Speech recognition with weighted finite-state transducers. *In Springer Handbook of Speech Processing*, pages 559–584. Springer.

[Nagao, 1984] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *In Proc. Of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.

[Nesterov, 1983] Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate o(1/k2).

[Neubig, 2017] Neubig, G. (2017). Neural machine translation and sequence-to-sequence models : A tutorial. *CoRR*, abs/1703.01619.

[Ng *et al.*, 2009] Ng, T., Nguyen, K., Zbib, R. et Nguyen, L. (2009). Improved morphological decomposition for arabic broadcast news transcription. *In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4309–4312. IEEE.

[Nguyen *et al.*, 2016] Nguyen, D., Doğruöz, A. S., Rosé, C. P. et de Jong, F. (2016). Computational sociolinguistics : A survey. *Computational Linguistics*, 42(3):537–593.

[Och, 1999] Och, F. J. (1999). An Efficient Method for Determining Bilingual Word Classes. *In Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, pages 71–76. Association for Computational Linguistics.

[Och, 2003] Och, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

[Och et Ney, 2003] Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

[O'Shaughnessy, 1988] O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE potentials*, 7(1):29–32.

[Papineni *et al.*, 2002] Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Parker *et al.*, 2011] Parker, R., Graff, D., Chen, K., Kong, J. et Maeda, K. (2011). Arabic gigaword fifth edition ldc2011t11. *Philadelphia : Linguistic Data Consortium.*

[Peddinti *et al.*, 2015] Peddinti, V., Povey, D. et Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *In Sixteenth Annual Conference of the International Speech Communication Association.*

[Poplack, 1980] Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol : toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

[Poplack et Walker, 2003] Poplack, S. et Walker, J. (2003). Pieter muysken, bilingual speech : a typology of code-mixing. cambridge : Cambridge university press, 2000. pp. xvi+306. 39:678 – 683.

[Povey *et al.*, 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. et Vesely, K. (2011). The kaldi speech recognition toolkit. *In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

[Povey et Woodland, 2002] Povey, D. et Woodland, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. *In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–105–I–108.

[Qian, 1999] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145 – 151.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rao *et al.*, 2015] Rao, K., Peng, F., Sak, H. et Beaufays, F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE.

[Rasooli et Collins, 2019] Rasooli, M. S. et Collins, M. (2019). Low-resource syntactic transfer with unsupervised source reordering. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3845–3856, Minneapolis, Minnesota. Association for Computational Linguistics.

[Rath *et al.*, 2013] Rath, S., Povey, D., Veselý, K. et Cernocky, J. (2013). Improved feature processing for deep neural networks. *Proc. Interspeech*.

[Rosenfeld, 1995] Rosenfeld, R. (1995). Optimizing lexical and n-gram coverage via judicious use of linguistic data. *In Fourth European Conference on Speech Communication and Technology*.

[Ruder, 2016] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747.

[Rumelhart *et al.*, 1986] Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[Sahraeian et Compernolle, 2016] Sahraeian, R. et Compernolle, D. V. (2016). Using weighted model averaging in distributed multilingual DNNs to improve low resource ASR. *Procedia Computer Science*, 81:152 – 158. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[Salloum et Habash, 2012] Salloum, W. et Habash, N. (2012). Elissa : A dialectal to standard arabic machine translation system. *In Proceedings of COLING 2012 : Demonstration Papers*, pages 385–392.

[Sarkar, 2016] Sarkar, K. (2016). Part-of-speech tagging for code-mixed indian social media text at icon 2015. *arXiv preprint arXiv :1601.01195*.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees.

[Schulz et Keller, 2016] Schulz, S. et Keller, M. (2016). Code-switching ubique est-language identification and part-of-speech tagging for historical mixed text. *In Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–51.

[Shaalan *et al.*, 2010] Shaalan, K., Aref, R. et Fahmy, A. (2010). An approach for analyzing and correcting spelling errors for non-native arabic learners. *In Informatics and Systems (INFOS), 2010 The 7th International Conference on*, pages 1–7. IEEE.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

[Sinha et Thakur, 2005] Sinha, R. M. K. et Thakur, A. (2005). Machine translation of bilingual hindi-english (hinglish) text.

[Smaïli *et al.*, 2018] Smaïli, K., Fohr, D., González-Gallardo, C., Grega, M., Janowski, L., Jouvet, D., Komorowski, A., Kozbial, A., Langlois, D., Leszczuk, M., Mella, O., Menacer, M. A., Mendez, A., Linhares Pontes, E., Sanjuan, E., Swist, D., Torres-Moreno, J.-M. et Garcia-Zapirain, B. (2018). A First Summarization System of a Video in a Target Language. *In MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems*, pages 1–12, Wroclaw, Poland.

[Smit *et al.*, 2018] Smit, P., Gangireddy, S., Enarvi, S., Virpioja, S. et Kurimo, M. (2018). Aalto system for the 2017 arabic multi-genre broadcast challenge. *In Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*, pages 338–345, United States. IEEE.

[Snover *et al.*, 2006] Snover, M., Dorr, B. J., Schwartz, R. et Micciulla, L. (2006). A study of translation edit rate with targeted human annotation.

[Snover *et al.*, 2009] Snover, M., Madnani, N., Dorr, B. et Schwartz, R. (2009). Fluency, adequacy, or HTER ? Exploring different human judgments with a tunable MT metric. *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

[Solorio et Liu, 2008] Solorio, T. et Liu, Y. (2008). Learning to predict code-switching points. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

[Soltau *et al.*, 2011] Soltau, H., Mangu, L. et Biadsy, F. (2011). From modern standard arabic to levantine ASR : Leveraging gale for dialects. *In 2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 266–271.

[Soltau *et al.*, 2009] Soltau, H., Saon, G., Kingsbury, B., Kuo, H.-K., Mangu, L., Povey, D. et Emami, A. (2009). Advances in arabic speech transcription at ibm under the darpa gale program. *IEEE Transactions on Audio, Speech and Language Processing*, 17:884–894.

[Soltau *et al.*, 2014] Soltau, H., Saon, G., Mangu, L., Kuo, H.-K., Kingsbury, B., Chu, S. et Biadsy, F. (2014). *Automatic Speech Recognition*, pages 409–459. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Somers, 1999] Somers, H. (1999). Example-based machine translation. *Machine Translation*, 14(2):113–157.

[Stolcke, 2000] Stolcke, A. (2000). Entropy-based pruning of backoff language models. *arXiv preprint cs/0006025*.

[Stolcke *et al.*, 2011] Stolcke, A., Zheng, J., Wang, W. et Abrash, V. (2011). Srilm at sixteen : Update and outlook. *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE SPS.

[Suontausta et Häkkinen, 2000] Suontausta, J. et Häkkinen, J. (2000). Decision tree based text-to-phoneme mapping for speech recognition. *In INTERSPEECH*.

[Sutskever *et al.*, 2014] Sutskever, I., Vinyals, O. et Le, Q. V. (2014). Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*, pages 3104–3112.

[Tachicart et Bouzoubaa, 2014] Tachicart, R. et Bouzoubaa, K. (2014). A hybrid approach to translate moroccan arabic dialect. *In 2014 9th International Conference on Intelligent Systems : Theories and Applications (SITA-14)*, pages 1–5.

[Taylor, 2005a] Taylor, P. (2005a). Hidden markov models for grapheme to phoneme conversion. pages 1973–1976.

[Taylor, 2005b] Taylor, P. (2005b). Hidden markov models for grapheme to phoneme conversion. *In Ninth European Conference on Speech Communication and Technology*.

[Tillmann *et al.*, 1997] Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. et Sawaf, H. (1997). Accelerated dp based search for statistical translation. *In Fifth European Conference on Speech Communication and Technology*.

[Uchida et Zhu, 2001] Uchida, H. et Zhu, M. (2001). The universal networking language beyond machine translation. *In International Symposium on Language in Cyberspace, Seoul*, pages 26–27.

[Udhyakumar *et al.*, 2004] Udhyakumar, N., Kumar, C., Srinivasan, R. et Swaminathan, R. (2004). Decision tree learning for automatic grapheme-to-phoneme conversion for tamil.

[van der Wees *et al.*, 2016] van der Wees, M., Bisazza, A. et Monz, C. (2016). A simple but effective approach to improve Arabizi-to-English statistical machine translation. *In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 43–50, Osaka, Japan. The COLING 2016 Organizing Committee.

[Venkataraman et Wang, 2003] Venkataraman, A. et Wang, W. (2003). Techniques for effective vocabulary selection. *CoRR*, cs.CL/0306022.

[Vergyri et Kirchhoff, 2004] Vergyri, D. et Kirchhoff, K. (2004). Automatic diacritization of arabic for acoustic modeling in speech recognition. *In Proceedings of the workshop on computational approaches to Arabic script-based languages*, pages 66–73. Association for Computational Linguistics.

[Veselý *et al.*, 2013] Veselý, K., Ghoshal, A., Burget, L. et Povey, D. (2013). Sequence-discriminative training of deep neural networks. *In INTERSPEECH*.

[Vincent *et al.*, 2008] Vincent, P., Larochelle, H., Bengio, Y. et Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *In Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

[Waibel *et al.*, 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. et Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339.

[Wang *et al.*, 2007] Wang, C., Collins, M. et Koehn, P. (2007). Chinese syntactic reordering for statistical machine translation. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745.

[Weaver, 1955] Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23.

[Weiss *et al.*, 2017] Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y. et Chen, Z. (2017). Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581.

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

[Wu *et al.*, 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. et Dean, J. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

[Xiang *et al.*, 2006] Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R. et Makhoul, J. (2006). Morphological decomposition for arabic broadcast news transcription. *In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

[Yaseen *et al.*, 2006] Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., Krauwer, S., Bendahman, C., Fersøe, H., Rashwan, M., Haddad, B., Mukbel, C., Mouradi, A., Al-Kufaishi, A., Shahin, M., Chenfour, N. et Ragheb, A. (2006). Building annotated written and spoken Arabic LRs in NEMLAR project. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

[Yolchuyeva *et al.*, 2019] Yolchuyeva, S., Németh, G. et Gyires-Tóth, B. (2019). Grapheme-to-phoneme conversion with convolutional neural networks. *Applied Sciences*, 9(6):1143.

[Young *et al.*, 1994] Young, S., Odell, J. et Woodland, P. (1994). Tree-based state tying for high accuracy modelling. *In HUMAN LANGUAGE TECHNOLOGY : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

[Zeiler, 2012] Zeiler, M. D. (2012). Adadelta : An adaptive learning rate method.

[Zens *et al.*, 2002] Zens, R., Och, F. J. et Ney, H. (2002). *Phrase-Based Statistical Machine Translation*, pages 18–32. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Zerrouki *et al.*, 2014] Zerrouki, T., Alhawaity, K. et Balla, A. (2014). Autocorrection of arabic common errors for large text corpus. *ANLP 2014*, page 127.

[Zhang *et al.*, 2014] Zhang, Q., Chen, H. et Huang, X. (2014). Chinese-english mixed text normalization. *In Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 433–442, New York, NY, USA. ACM.

[Zhang et Yang, 2017] Zhang, Y. et Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv :1707.08114*.

Résumé

Les travaux de recherche ont été développés dans le cadre du projet AMIS (*Access to Multilingual Information and opinionS*). AMIS est un projet européen dont l'objectif principal est d'aider les personnes à comprendre l'idée générale d'une vidéo dans une langue étrangère en générant un résumé automatique de cette dernière dans une langue compréhensible par l'utilisateur. Concrètement, répondre à ce besoin engage une interaction judicieuse entre plusieurs modules impliquant plusieurs défis scientifiques, à savoir : l'extraction automatique du résumé d'une vidéo et/ou d'un texte, la reconnaissance automatique de la parole et la traduction automatique. Dans le cadre de cette thèse, nous nous sommes concentrés sur la reconnaissance et la traduction automatique de la parole de vidéos arabes et dialectales.

Les approches statistiques proposées dans la littérature pour la reconnaissance automatique de la parole sont indépendantes de la langue et elles sont applicables à l'arabe standard. Cependant, cette dernière présente quelques caractéristiques que nous devons prendre en considération afin de booster les performances du système de reconnaissance. Parmi ces caractéristiques on peut citer l'absence de l'indication des voyelles courtes dans le texte ce qui rend difficile leur apprentissage par le modèle acoustique. On trouve aussi la simplification de l'écriture de certains mots en remplaçant une lettre par une autre ce qui impacte directement le modèle de langage. En partant de ces caractéristiques, nous avons proposé plusieurs approches de modélisation acoustique et/ou de langage afin de mieux reconnaître la parole arabe.

Dans le monde arabe, l'arabe standard n'est pas la langue maternelle, elle est apprise à l'école, c'est pourquoi dans les conversations quotidiennes, on utilise le dialecte, un arabe inspiré de l'arabe standard, mais pas seulement. Nous avons travaillé sur l'adaptation du système développé pour l'arabe standard au dialecte algérien qui est l'une des variantes de la langue arabe les plus difficiles à reconnaître par les systèmes de reconnaissance automatique de la parole. Cela est dû aux mots empruntés d'autres langues, au *code-switching* (phénomène qui se produit lorsqu'un locuteur alterne entre deux ou plusieurs langues dans ses discours) et au manque de ressources nécessaires pour l'apprentissage des différents modèles. Notre proposition pour remédier à ces problèmes est de tirer profit des données orales et textuelles d'autres langues impactant le dialecte pour entraîner les modèles nécessaires à la reconnaissance du dialecte.

Le texte résultant de la reconnaissance de la parole arabe a alors été utilisé pour la traduction automatique. Nous avons réalisé dans un premier temps une étude comparative entre l'approche statistique à base de segments et l'approche neuronale utilisées dans le cadre de la traduction automatique. Ensuite, nous nous sommes intéressés à la traduction du texte *code-switché* généralement considéré comme du bruit lors de l'apprentissage et/ou du test. Notre étude portait sur le mélange de l'arabe et de l'anglais dans un corpus parallèle extrait de documents officiels des nations unies. Nous avons construit un corpus parallèle où les phrases sources sont des phrases *code-switchées* arabe-anglais et les phrases cibles sont leurs traductions dans les deux langues arabe et anglaise. En se basant sur cette ressource, nous avons travaillé sur l'adaptation des systèmes de traduction automatique afin d'améliorer la traduction du texte *code-switché*.

Le système séquentiel proposé dans le cadre du projet AMIS pour la traduction de vidéos soulève plusieurs problèmes dus à la propagation des erreurs. Afin de minimiser ces erreurs, nous avons travaillé sur l'adaptation du vocabulaire du système de reconnaissance automatique de la parole et sur la proposition d'une nouvelle modélisation transformant directement un signal de

la parole dans une langue A en une séquence de mots dans une autre langue B.

Mots-clés: reconnaissance automatique de la parole, traduction automatique, arabe standard, dialecte algérien, *code-switching*.

Abstract

This research has been developed in the framework of the project AMIS (Access to Multilingual Information and opinionS). AMIS is an European project which aims to help people to understand the main idea of a video in a foreign language by generating an automatic summary of it. In concrete terms, meeting this need involves a judicious interaction between several modules involving several scientific challenges, namely : automatic summarization of a video and/or a text, automatic speech recognition and machine translation. In this thesis, we focus on the automatic recognition and translation of the speech of Arabic and dialectal videos.

The statistical approaches proposed in the literature for automatic speech recognition are language independent and they are applicable to modern standard Arabic. However, this language presents some characteristics that we need to take into consideration in order to boost the performance of the speech recognition system. Among these characteristics we can mention the absence of short vowels in the text, which makes their training by the acoustic model difficult. We also mention the simplification of the writing of certain words by replacing one letter by another, which directly impacts the language model. Based on these characteristics, we proposed several approaches to acoustic and/or language modeling in order to better recognize the Arabic speech.

In the Arab world, modern standard Arabic is not the mother tongue, it is taught at school, that is why daily conversations are carried out with dialect, an Arabic inspired from modern standard Arabic, but not only. We worked on the adaptation of the speech recognition system developed for the modern standard Arabic to the Algerian dialect, which is one of the most difficult variants of the Arabic language to recognize by automatic speech recognition systems. This is mainly due to the borrowed words from other languages, the code-switching (phenomenon that occurs when a speaker alternates between two or more languages within an utterance or discourse) and the lack of resources to train the different models. Our approach to overcome all these problems is to take advantage from oral and textual data of other languages that have an impact on the dialect in order to train the required models for dialect speech recognition.

The resulting text from Arabic speech recognition system was then used for machine translation. As a starting point, we conducted a comparative study between the phrase based approach and the neural approach used in machine translation. Then, we focused on the translation of the code-switched text that is generally considered as noise in the training and/or testing stages. Our study focused on the mix of Arabic and English in a parallel corpus extracted from official documents of the United Nations. We constructed a parallel corpus where the source sentences are Arabic-English code-switched sentences and the target sentences are their translations in both Arabic and English. Based on this resource, we worked on the adaptation of machine translation systems in order to improve the translation of the code-switched text.

The pipline system proposed in the AMIS project for video translation raises several issues due to the error propagation. In order to minimize these errors, we worked on the adaptation of the vocabulary of the automatic speech recognition system and on the proposition of a new model that directly transforms a speech signal in language A into a sequence of words in another language B.

Keywords: automatic speech recognition, machine translation, modern standard Arabic, Algerian dialect, *code-switching*.