Approche hybride pour la reconnaissance automatique de la parole en langue arabe

Abir Masmoudi Dammak

Université du Maine

Université de Sfax

Thèse

Pour l'obtention du titre de docteur en :

Informatique

Approche hybride pour la reconnaissance automatique de la parole pour la langue arabe

préparée par

Abir MASMOUDI DAMMAK

Soutenue le 21 Septembre 2016 devant

Prof. Rim ZITOUNE FAIZ	IHEC-Université de Carthage	Présidente
Prof. Frederic BECHET	Université d'Aix-Marseille	Rapporteur
Prof. Rim ZITOUNE FAIZ	IHEC-Université de Carthage	Rapporteur
Prof. Lamia BELGUITH	FSEGS-Université de Sfax	Directeur de thèse
Prof. Yannick ESTEVE	Université du Maine	Directeur de thèse

Abstract :

Récemment, les tentatives dans le domaine de la RAP pour la langue arabe ont éveillé l'attention de quelques chercheurs. En fait, la majorité de ces tentatives ont mis l'accent sur la norme officielle de la langue arabe qui est connu comme l'arabe moderne standard (MSA). En revanche, le MSA ne présente pas la langue des communications courantes dans tous les pays arabes. De ce fait, on conçoit l'existence de plusieurs variétés arabes utilisées dans la vie quotidienne pour la communication ordinaire des communautés. Certes, ces différents dialectes arabes possèdent une forme parlée et non écrite et se distinguent par des caractéristiques phonologiques, morphologiques, syntaxiques et lexicales importantes qui se diffèrent d'un dialecte à un autre et même avec le MSA.

En règle générale, la nature statistique des approches exige de disposer d'une grande quantité de ressources à savoir, grands corpus de texte, grands corpus de parole, dictionnaires de prononciation pour le développement d'un SRAP. Néanmoins, ces ressources ne sont pas disponibles directement pour des dialectes arabes. De ce fait, le développement des SRAP pour les dialectes arabes se heurte à de multiples difficultés de différentes natures auxquelles elle doit faire face. Ces difficultés peuvent être résumées comme suit : *i)* l'absence ou quasi-absence de grandes quantités de ressources disponibles pour ces dialectes. *ii)* les caractéristiques phonétiques et phonologiques de cette langue ne sont pas encore bien connues. *iii)* l'absence d'une orthographe standard vu l'absence d'académies pour ces dialectes.

Dans cette perspective, les travaux de cette thèse s'intègrent dans le cadre du développement d'un SRAP pour le dialecte tunisien. Une première partie des contributions consiste à développer une variante de CODA (Conventional Orthography for Arabic Dialectal) pour le dialecte tunisien. En fait, cette convention est conçue dans le but de fournir une description détaillée des directives appliquées au dialecte tunisien. Compte tenu des lignes directives de CODA, nous avons constitué nos ressources dialectales utilisées par la suite pour le développement du SRAP. Ainsi, nous avons collecté deux types de corpus : *i)*corpus textuel et de paroles nommé TARIC : Corpus de l'interaction des chemins de fer de l'arabe tunisien dans le domaine de la SNCFT. En effet, ce corpus est constitué des enregistrements audio et la transcription manuelle de ces enregistrements. *ii)* Des données textuelles recueillies à partir des sites Web notamment les blogs Tunisiens et de la translittération des données dialectales écrites en caractère latin.

Outre ces ressources notamment les corpus de texte et de parole, le dictionnaire de prononciation s'impose d'une manière indispensable pour le développement d'un SRAP. À ce propos, dans la deuxième partie des contributions, nous visons la création d'un système nommé conversion G2P qui permet de générer automatiquement ce dictionnaire phonétique. Ainsi, deux types d'approches vont être utilisés. L'approche à base de règles. Le principe de cette approche est d'utiliser des règles phonétiques pour la conversion G2P. L'approche guidée par les données. En fait, nous avons utilisé une approche pour la conversion G2P basé sur une méthode probabiliste : Conditional Random Fields (CRF).

Toutes ces ressources décrites avant sont utilisées pour adapter un SRAP pour le MSA du laboratoire LIUM au dialecte tunisien dans le domaine de la SNCFT. L'évaluation de notre système donné lieu WER de 22,6% sur l'ensemble de test.

Key words :
Dialecte tunisien, reconnaissance automatique de la parole, ressources écrites et orales, conversion G2P, approche à base de règles, approche probabiliste, CRF.

Remerciement

Mes respectueux remerciements s'adressent à Mme. Lamia HADRICH BELGUITH professeur à la faculté des sciences économiques et de gestion de Sfax et Mr. Yannick ESTEVE professeur à l'université du Maine d'avoir accepté d'être mes directeurs de thèse et de m'avoir fait profiter de leurs expériences respectives. Je tiens à les remercier vivement pour les conseils fructueux qu'ils m'ont prodigués, pour leurs orientations ciblées, pour leur disponibilité, pour la confiance qu'ils ont su m'accorder pour leur rigueur et leur compréhension tout au long de ces années de travail.

Je tiens aussi à exprimer ma profonde gratitude envers Mme. Mariem ELLOUZE et Mr. Fethi BOUGARES mes encadrants de thèse pour ses conseils, ses aides et ses encouragements, ainsi que l'intérêt continu qu'ils ont porté à mon travail.

Je remercie le président de jury Mme. Rim ZITOUNI FAIEZ, Professeur à l'Institut des hautes études commerciales de Carthage pour l'honneur qu'elle m'accorde en jugeant ce travail. Ainsi, je remercie vivement mes rapporteurs Mme. Rim ZITOUNI FAIEZ et Mr. Frédéric BÉCHET qui ont accepté d'évaluer le présent travail.

Je voudrais exprimer mes remerciements à tous les membres de l'équipe ANLP-RG ainsi que les membres de l'équipe LIUM qui m'ont aidée par leurs conseils et leurs critiques.

Je remercie vivement Mr. Nizar HABASH professeur à l'université de New York à Abu Dhabi avec qui j'ai eu l'occasion de travailler et qui ont contribué aux recherches présentées dans ce manuscrit.

Une pensée toute spéciale à ma famille qui m'a apporté le soutien dont j'avais besoin pour mener à bien ce travail. Je remercie mes parents pour leur confiance et pour leurs encouragements. Ma gratitude s'adresse également à mes frères, à ma soeur Mariem et ma cousine Marwa.

Enfin, mes derniers remerciements s'adressent à mon mari Ahmed DAMMAK que je ne cesserai de remercier. Son aide, son amour et ses encouragements durant ces années de thèse m'ont été des plus précieux.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Depuis des années, la Reconnaissance Automatique de la Parole (RAP) est un domaine de la science ayant toujours eu un grand attrait auprès des chercheurs comme auprès du grand public. Ce domaine vise par le biais de ses systèmes, à « décoder » un signal vocal acoustique en une chaîne de mots. Ainsi, l'une des majeures applications de la RAP est l'Interaction Homme-Machine(IHM). À ses balbutiements, les projections sur ses applications étaient très optimistes : quoi de plus naturel que de parler à une machine, sans avoir à s'encombrer d'un clavier ou d'une souris ? Les applications qu'ont peut imaginer sont nombreuses : les serveurs vocaux, les réservations des vols, l'apprentissage de langues, etc. Nous citons aussi les commandes vocales des machines ou des robots, la saisie vocale de données, l'aide aux handicapés (contrôle par voix, machines à parler vocale), l'utilisation de la reconnaissance de la parole dans les jeux électroniques.

Les premières expériences de développement des systèmes de RAP ont été perçues pour plusieurs langues telles que la langue anglaise, française aussi bien les langues asiatiques. Malgré que la langue arabe soit très répandue dans le monde, les recherches fructueuses effectuées dans le domaine de la RAP pour l'arabe restent très limitées en comparaison avec d'autres langues de même rang comme le Chinois. Ce qui explique le manque du support pour la langue arabe dans la majorité des applications IHM.

Récemment, les tentatives dans le domaine de la RAP pour la langue arabe ont éveillé l'attention de quelques chercheurs. En fait, la majorité de ces tentatives ont mis l'accent sur la norme officielle de la langue arabe qui est connue comme l'Arabe Moderne Standard (MSA). En revanche, MSA ne présente pas la langue des communications courantes dans les pays arabes. De ce fait, en contemplant de prés, on conçoit l'existence de plusieurs variétés arabes le levantine, l'égyptien, l'algérien, le marocain, le tunisien, etc... considéres comme des dialectes arabes dérivées de MSA et utilisées dans la vie quotidienne pour la communication ordinaires des communautés. Cette variété de dialectes que nous pouvons même trouver au sein du même pays, évoque un problème majeur dans le traitement automatique de la langue arabe. Certes, ces différents dialectes arabes possèdent une forme parlée et non écrite et se distinguent par des caractéristiques phonologiques, morpholo-

giques, syntaxiques et lexicales importantes qui se diffèrent d'un dialecte à un autre et même avec la forme standard de la langue arabe. Cette situation est dénommée « *diglossie* », ce terme est inventé par [Fishman 1967] qui signifie la situation où il existe en usage deux langues apparentées génétiquement et structurellement et dont les distributions fonctionnelles sont complémentaires.

Comme premier pas dans cette direction, nous avons choisi l'arabe dialectal tunisien comme étant un exemple de l'arabe du Maghreb et qui est encore peu étudié du point de vue du traitement automatique.

En règle générale, le développement d'un Système de RAP (SRAP) pour une langue spécifique exige en premier lieu la construction d'un corpus de parole en grande quantité. Ce corpus doit être transcrit de façon orthographique et phonétique. Un nombre important de différents locuteurs est également nécessaire en vue de modéliser des accents différents et par conséquent aboutir à la création d'un SRAP indépendant d'un locuteur donné [Rabiner 1989]. De plus, des données textuelles pour l'apprentissage des modèles de langage du système doivent être requises.

Le dialecte tunisien souffre d'un manque de données linguistiques contrairement au MSA qui se caractérise par l'existence d'une grande quantité de ressources pour la création d'un SRAP. De ce fait, le développement d'un SRAP fiable pour le dialecte tunisien représente une tâche cruciale et assez difficile. Ceci est principalement dû à la nature diglossique du dialecte tunisien aussi qu'aux difficultés d'estimation de la transcription phonétique. Certainement, le développement d'un SRAP pour le dialecte tunisien souffre encore de nombreux problèmes reliés aux différents facteurs de variabilité qui peuvent être résumés comme suit :

- En premier abord, la quantité de ressources disponible est très limitée à cause de l'existence de nombreux dialectes arabes dans le monde, voir même des divergences entre les dialectes au sein du même pays.
- En second lieu, le dialecte tunisien ne dispose pas d'une orthographe standard vue l'absence d'académies pour ce dialecte et comme nous l'avons déjà indiqué auparavant, il est primordialement parlé et non écrit. De ce fait, la tâche de collection des ressources notamment de corpus textuels est d'autant plus difficile qu'il n'y a pas de conventions de transcription admises par la communauté scientifique.
- En troisième lieu, il existe encore des difficultés au niveau de la transcription phonétique, puisque les caractéristiques phonétiques et phonologiques ne sont pas encore bien connues. D'une façon générale, le dialecte tunisien se caractérise par l'ajout d'autres phonèmes supplémentaires qui ne peuvent pas être estimés directement ni par le script, ni par les signes diacritiques. Compte tenu de ce qui précède, les techniques de transcriptions phonétiques

dédiées pour le MSA ne sont pas utilisables directement pour ce dialecte sans faire des modifications.

- En dernier lieu et dans le même ordre d'idées, la transcription phonétique du dialecte tunisien est confrontée à de nombreux problèmes notamment les problèmes morpho-phonémiques, l'élision et également les variations phonologiques et phonétiques de certaines consonnes. De plus, cette transcription doit tenir en compte l'apparition de nouveaux phénomènes comme l'assimilation et métathèses. Par ailleurs, la transcription phonétique doit trouver des solutions dans le cas où il existe des mots étrangers dans la langue et des irrégularités d'orthographe. À ces problèmes s'ajoutent d'autres spécificités au dialecte tunisien dont nous citons l'absence de voyelles courtes dans les mots. En effet, une ambigüité dans la lecture voir même la sémantique du mots est engendrée. Cette ambigüité apparait notamment dans la détermination de la forme phonétique d'un mot qui est devient plus difficile par rapport au mot voyellé.

Dans cette perspective, les travaux de cette thèse s'intègrent dans le cadre du développement d'un SRAP pour le dialecte tunisien, dans le domaine de renseignement ferroviaire. Ce dialecte est conçu en tant qu'une langue peu dotée qui ne dispose pas suffisamment de ressources linguistiques en quantité et en qualité pour la construction d'un SRAP. Cette situation semble plus critique vu que la mise en place d'une telle quantité de ressources nécessite un effort supplémentaire important. De ce fait, afin de construire un SRAP pour le dialecte tunisien, nous avons besoin de deux types de corpus à savoir, des enregistrements audio et des textes écrits correspondants.

En outre, dans notre travail nous opté pour la création de nos propres corpus audio et textuels nommée TARIC : Corpus de l'interaction des chemins de fer de l'arabe tunisien dans le domaine ferroviaire de la SNCFT [1] : **S**ociété **N**ationale de **C**hemins de **F**er **T**unisiens. Ce corpus audio représente des conversations entre des passagers et des agents de la SNCFT. Le but de ces conversations consiste à demander en dialecte tunisien des informations sur les services de chemin de fer dans une gare ferroviaire. Ces demandes correspondent aux types de train, ses horaires, sa destination, le prix et la réservation des billets.

Par ailleurs, jusqu'à maintenant la transcription automatique du dialecte tunisien se heurte au manque d'outils et de normes d'écritures. De surcroît, la transcription manuelle semble une tâche pénible et couteuse ceci à part l'absence de conventions de transcription admises par la communauté scientifique. Ainsi, avant d'entamer la transcription, nous avons développé une convention d'écriture nommée CODA (**C**onventional **O**rthographique for **D**ialectal **A**rabic) au profit du

1. http ://www.sncft.com.tn/

dialecte tunisien. Le but de CODA est d'obtenir des données cohérentes et consistantes et que chaque mot aura une seule représentation orthographique.

Dans le but de pallier le problème de la carence des données, nous avons essentiellement considéré des ressources issues à partir de deux méthodes de collection de données de grande quantité et de façon rapide. Premièrement, une attention particulière a été notamment apportée à une approche intéressante qui vise à « aspirer » les sites Web en dialecte tunisien et à filtrer ensuite les données récupérées pour les rendre exploitables. Dans cette perspective, nous avons choisi les sites de blogs écrits en dialecte tunisien (écrit en alphabet arabe). Deuxièmement, nous avons fait recours à l'utilisation des formes de communication écrites (i.e. email, chat, SMS, commentaires,...) en tant qu'un point de départ pour la construction de grands corpus de façon automatique. Néanmoins, la majorité de ces messages et commentaires sont écrits avec l'alphabet latin. Pour ces motifs, nous avons développé un outil de translittération pour retranscrire les données recueillies en caractères arabe. Ainsi, il faut s'assurer que cette conversion est effectuée selon la convention de l'orthographe CODA de l'arabe dialectal. De ce fait, la totalité de ces données sera utilisée pour apprendre et améliorer les modèles de langage.

En sus des ressources nécessaires notamment les corpus de texte et de parole, le dictionnaire de prononciation s'impose pour le développement d'un SRAP. Ce dictionnaire représente la relation entre la modélisation acoustique et la modélisation linguistique. Conformément à ce qui précède, un des objets d'intérêt de notre travail consiste à créer un dictionnaire de prononciation pour le dialecte tunisien. À cet égard, notre proposition opère par l'utilisation d'une approche hybride qui combine une première conversion Graphème en Phonème purement symbolique avec une deuxième conversion purement probabiliste. Pour illustrer cet aspect hybride, nous proposons d'utiliser en premier lieu l'approche à base de règles pour la construction de ce dictionnaire. En deuxième lieu, nous utilisons une méthode statistique pour améliorer notre outil de phonétisation. Cette deuxième méthode est basée sur CRF «Conditional Random Fields».

À l'issue de ces traitements, nous utiliserons ces ressources sont utilisées pour adapter un SRAP pour le MSA développé au sein de du laboratoire LIUM [2], au dialecte tunisien dans le domaine de ferroviaire.

Structure du document

Cette thèse est organisée comme suit :

Le premier chapitre de ce manuscrit s'intéresse les caractéristiques générales de la langue arabe : du littéral au dialectal. Nous décrivons les différentes registres

2. http ://www-lium.univ-lemans.fr/fr/content/bienvenue

linguistiques de la langue arabe ainsi que le passage historique de l'arabe classique vers le MSA jusqu'à l'arrivée à l'arabe dialectal. Le dialecte tunisien a été sélectionné comme un exemple de l'arabe dialectal. Nous présentons d'une part la situation linguistique de ce dialecte et nous faisons d'autre part la comparaison du dialecte tunisien du nord avec celui de sud sur différents niveaux : ***i)*** le niveau morphologique, ***ii)*** le niveau phonologique et ***iii)*** le niveau lexical. Nous clôturons ce chapitre par citer les principales différences du dialecte tunisien par rapport au MSA.

Le deuxième chapitre est composé en deux principales parties. La première partie est consacrée à la présentation de l'état de l'art dans le domaine de la reconnaissance automatique de la parole. Dans cette partie, nous détaillons le principe de fonctionnement d'un SRAP en mettant l'accent sur les trois composantes principales à savoir, le module acoustique, le module de modélisation du langage et le dictionnaire de prononciation. Dans la deuxième partie, nous présentons un aperçu sur quelques SRAP développés dans la littérature pour les langues peu dotées, allant de l'acquisition des ressources textuelles et de parole jusqu'aux expérimentations.

Nous détaillons, ensuite, dans le troisième chapitre les différentes approches théoriques existantes dans la littérature pour la transcription phonétique qu'on appelle aussi la conversion graphème en phonème (G2P). Ces approches sont classées en trois catégories : l'approche manuelle, l'approche à base de règles et l'approche guidée par les données. Ce chapitre est pour nous l'occasion de présenter les approches principales et les techniques existantes dans la littérature pour la conversion G2P tout en discutant les points forts et les faiblesses de ces approches.

Dans le quatrième chapitre, nous présentons une vue globale des lignes directives de notre convention de normalisation CODA. À cette fin, nous commençons par exposer les objectifs et les principes de CODA. À cet égard, la première partie de ce chapitre est clôturé par une présentation détaillée des lignes directives de CODA. Compte tenu de ces lignes directives, nous allons recueillir nos corpus qui vont être utilisés pour le développement d'un SRAP pour le dialecte tunisien. Ces données recueillies comportent d'une part des signaux de parole, et d'autre part des données textuelles. En définitive, la deuxième partie de ce chapitre décrit notre méthode pour la construction des ressources.

Le cinquième chapitre est dédié à une étude détaillée sur la conversion G2P du dialecte tunisien. Ainsi, nous avons commencé ce chapitre par un survol sur les problèmes de cette conversion. Nous visons dans ce chapitre de s'intéresser plus aux solutions proposées pour résoudre les problèmes et les règles utilisées pour la tâche de conversion G2P de cette langue en utilisant notre approche hybride : symbolique et statistique.

Le chapitre six a pour objet la description de nos résultats d'évaluation du premier SRAP pour le dialecte tunisien dans le domaine de renseignement ferroviaire.

Nous terminons ce manuscrit par une conclusion et quelques perspectives concernant les travaux futurs.

Chapitre 1

Caractéristiques générales de la langue arabe : du littéral au dialectal

Sommaire

1.1 Introduction

La langue arabe appartient à la famille des langues sémitiques, elle est utilisée comme vecteur de transmission religieux pour tous les croyants musulmans au nombre de 1 milliard et demi à travers les cinq continents du globe. Cette langue

a un statut spécial en tant que norme officielle du monde arabe. À cet effet, et sans contredit, la langue arabe est l'idiome qui a envahi la plus grande étendue des pays du monde entier puisqu'elle est la langue officielle de plus de 22 pays. Ainsi, elle est classée comme la 6 ème langue la plus parlée en fonction du nombre de locuteurs [Elmahdy 2012] selon l'Organisation des Nations Unies. En effet, elle constitue un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial. Autrefois, son histoire s'est heurtée à divers évènements : conquêtes arabes, essor scientifique, colonisation occidentale, tentative de réforme grammaticale ou adaptation au monde moderne. Ces derniers années elle a pris encore l'essor suite aux révolutions dans divers pays arabes comme la Tunisie, l'Égypte, la Syrie ; etc.

C'est à dessein que ce chapitre a pour but de décrire l'historique de la langue arabe tout en indiquant sa particularité et sa répartition géographique actuelle. Encore, nous présenterons les registres linguistiques de l'arabe ainsi que le passage historique de l'arabe classique vers le MSA jusqu'à l'arrivée à l'arabe dialectal. Aussi, nous exposerons la différence entre la langue arabe et ses divers dialectes. Ensuite, nous allons spécifier les principales caractéristiques du dialecte tunisien qui est la langue traitée dans le cadre de ce travail. Nous allons clôturer ce chapitre par citer les principales différences du dialecte tunisien par rapport au MSA sur des différents niveaux à savoir, le niveau phonologique, morphologique, lexical et syntaxique.

1.2 La langue arabe

1.2.1 Historique de la langue arabe

Il est difficile d'aborder l'étude d'une langue sans faire référence à l'histoire qu'elle a vécue. Il en est ainsi pour l'arabe qui est une langue originaire de la péninsule Arabique et qui a connu une longue tradition orale avant d'être consignée à l'écrit [Sayah 2009].

À l'origine, au niveau de la péninsule arabe, la langue arabe appartient à la famille des langues sémitiques comme l'akkadien, l'hébreu, l'araméen et son expansion a touché même l'Afrique du nord et l'Asie mineur. Postérieurement, bien avant le VIe siècle de l'ère chrétienne, la littérature préislamique est représentée avant tout par la poésie jusqu'à l'apparition de l'islam. Avec la prédication du prophète « Mohamed [1] » et l'avancement de l'islam, porté d'abord par une conquête militaire, la langue dans laquelle s'est faite la Révélation consignée dans le Coran est née.

1. http ://en.wikipedia.org/wiki/Muhammad

Initialement, la langue arabe était limitée à la péninsule arabique. C'est autour du VIIe siècle et grâce à l'avènement de l'Islam et plus tard les conquêtes islamiques, la langue arabe a connu une grande expansion. Ce qui nous emmène que le développement de l'Islam au Xe siècle a permis l'arabe, en tant que langue religieuse, de se développer considérablement dans le monde musulman qui s'étend dans tout le Nord de l'Afrique et l'ensemble du Moyen-Orient. Assez tôt, comme de plus en plus les non-arabophones se convertirent à l'islam, le Coran devint le lien le plus important entre les musulmans, arabes et non-arabes, vénéré pour son contenu et admiré pour la beauté de son langage. En outre, les arabes, indépendamment de leur religion, et quelque soit l'origine ethnique des musulmans, tiennent de plus en plus à la langue arabe et la considèrent comme une norme idéale, devant la profonde évolution que les nouveaux usages sociaux et son histoire lui imposaient. Le grand rapport entre le Coran et l'arabe a donné à la langue un statut spécial qui a contribué à l'arabisation de populations diverses.

À travers des siècles, c'est la langue arabe qui a permis aux parlers natifs des pays arabes de se communiquer et de partager leurs cultures à travers le monde. Surtout, lors de l'avènement de l'Islam, elle est devenue la langue sacrée du Coran en exerçant des influences irrésistibles sur les peuples pour convertir à cette nouvelle religion. De plus, la langue arabe a recueilli des progrès étourdissants dans des domaines divers tels que la culture, la science grâce à la l'expansion territoriale de l'empire musulmane qui a fait de cette langue, une langue d'administration et de rédaction de manuscrits et de livres. Ainsi, il faut noter que le passage de l'arabe classique qui est ciblé en tant qu'une langue du Coran à l'arabe standard moderne (MSA) était fait à travers l'existence de la diversité au niveau des populations arabophones et ces cultures à travers des siècles. À son tour, le MSA représentant la langue officielle utilisée dans les communautés et la presse, a été influencé par des spécificités historiques et culturelles des populations appartenant au monde arabe en donnant naissance à l'arabe dialectal.

1.2.2 Particularités de la langue arabe

1.2.2.1 Absence des voyelles

La langue arabe est une langue sémitique qui s'écrit et se lit de droite à gauche. Il existe deux types de symboles dans l'alphabet arabe pour écrire des mots : les lettres et les signes diacritiques. Un des traits particuliers du système d'écriture arabe, par rapport aux langues latines, est la non distinction entre lettres minuscules et majuscules. Les lettres arabes, correspondant au 28 sons consonantiques arabes. Chaque lettre peut apparaître dans un maximum de quatre formes différentes, selon qu'elle se produit au début, au milieu ou à la fin d'un mot, ou en isolée. Les lettres sont principalement connectées. Pour des raisons phonétiques, les lettres de l'alphabet arabe sont classées en deux groupes : les lettres lunaires

et les lettres solaires.

Le second type de symboles dans l'alphabet arabe est les signes diacritiques. Dans l'écriture arabe il existe trois types de signes diacritiques : voyelles, nunation et shadda.

- **Les voyelles**, au nombre de trois, sont appelés aussi voyelles courtes : la damma ضمّة /Dammp/ [u] se représente comme une virgule (,) qui apparaissent sur le dessus d'une consonne, la fatha فتحة /fatHap/ [a] se représente comme un trait d'union (-) qui apparaît sur le dessus d'une consonne et la kasra كسرة /karsap/ [i] se représente comme un trait d'union (-) qui apparaissent en dessous d'une consonne.
- **Nunation** التنوين /Altnwyn/ peut seulement survenir dans la position finale d'un mot dans les nominales (noms, adjectifs et adverbes), où ils indiquent l'indétermination. Ils représentent la combinaison d'une voyelle courte et le marqueur non écrit /n/.
- **Shadda** الشَدّة /Al$dap/ est un signe diacritique ressemblant à la lettre minuscule [w]. Elle sert principalement à indiquer qu'une consonne est géminée, ce qui est l'équivalent d'un doublement de consonne. Elle est placée au dessus de la consonne en question. Elle est aussi employée dans les textes où les diacritiques sont absents pour limiter l'ambiguïté.

Dans la langue arabe, les lettres sont toujours écrites, les signes diacritiques sont facultatifs : l'écriture arabe peut être totalement voyellé, partiellement voyellé ou entièrement voyellé. L'absence de voyelles (la non-voyellation) dans les textes arabes génère plusieurs cas d'ambiguïtés et des problèmes lors de l'analyse automatique. En effet, l'ambiguïté grammaticale augmente si le mot est non voyellé. Cela est dû au fait qu'un mot non voyellé possède plusieurs voyellations possibles, et pour chaque voyellation est associée une liste différente de catégories grammaticales [Belguith 1999].

1.2.2.2 Agglutination

Contrairement aux langues latines, en arabe, « les articles[2] », « les prépositions[3] », « les pronoms[4] », etc. collent aux adjectifs, noms, verbes et particules

2. Les articles : par exemple ال.

3. Les prépositions sont : في, مع , لن , حتَّى , من, أيلَى, ل, ك, ب

4. Le pronom personnel en arabe est isolé ou affixé. Isolé, il correspond en français à : moi, toi, etc.

auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française [Souissi 1997]. Exemple : le mot arabe أَتَتذكرونَا /atat*krwnA/ correspond en Français à la phrase « Est ce que vous vous souvenez de nous ? ».

Cette caractéristique peut engendrer une ambiguïté au niveau morphologique. En effet, il est parfois difficile de distinguer entre « une proclitique [5] » ou « enclitique [6] » et un caractère original du mot. Par exemple, le caractère و dans le mot وَصَلَ /w/ [il est arrivé] est un caractère original alors que dans le mot وَفَتَحَ /wfataHa/ [et il a ouvert], il s'agit d'une proclitique [Belguith 2006].

1.2.3 Registres linguistiques de la langue arabe

En observant les périodes les plus importantes dans l'historique de la langue arabe, nous avons recours à découvrir trois registres linguistiques que nous allons citer dans la partie qui suit.

- **Arabe littéraire ancien ou classique** : Cette appellation désigne la langue arabe dans sa forme la plus classique et la plus ancienne. Cela concerne essentiellement tout le patrimoine culturel médiéval parvenu par écrit : le texte coranique, la poésie ancienne, la philosophie, l'historie, etc. La nature et l'origine de cette langue de la littérature antéislamique ont donné lieu à une évolution qui a abouti à l'apparition d'un arabe dit moderne ou standard.
- **Arabe moderne standard** : D'une manière générale, l'arabe standard ou l'arabe contemporain est le résultat de l'interaction entre l'arabe classique et les dialectes [Ammar 2012]. Dans le monde arabe, l'arabe moderne standard (MSA) est la langue des médias, de la vie intellectuelle et de la littérature. En outre, il représente la forme de l'arabe universel enseignée dans les écoles du monde arabe et même utilisée à des conférences et des discussions formelles.
- **L'arabe dialectal** : L'arabe dialectal est une forme extrêmement simplifiée de l'arabe classique et de l'arabe moderne. C'est la langue maternelle de chaque locuteur arabophone. Il est parlé dans tous les jours et qui ne s'embarrasse pas de toutes les règles rigides de la langue écrite et savante et qui évolue de plus en plus en fonction de l'époque et des besoins de communication. Il existe plusieurs dialectes arabes et ces formes linguistiques se diffèrent parfois d'une région à une autre même légèrement d'une ville à une autre. Principalement, nous distinguons que le monde arabe est divisé en deux aires

5. Les proclitiques représentent des conjonctions mono-consonnes ل ،و, des prépositions ل ،ب, un préverbe س indiquant le futur, un article ال qui permet la détermination d'un nom, etc

6. Les enclitiques sont les compléments de pronom ك, كمَا, هم, ه

dialectales : le groupe occidental et le groupe oriental. Ainsi la frontière naturelle entre ces deux groupes est marquée par le plus long fleuve du monde le « Nil». D'une part, le groupe occidental correspond aux variétés parlées en Égypte, à Djibouti, au Soudan, au Tchad, dans les États dits du Machrek (Irak, Syrie, Liban, Jordanie, Palestine et Koweït) et ceux des états de la péninsule Arabique (Arabie Saoudite, Yémen, Oman, Qatar, Émirats arabes unis, Koweït et Bahreïn). De sa part, le groupe maghrébin correspond aux variétés d'arabe parlées dans les pays du Maghreb (Tunisie, Algérie, Maroc, Libye, Mauritanie et Sahara occidental) en Andalousie (Espagne), ainsi que dans l'île de Malte.

1.2.4 Les différences entre la langue arabe et ses dialectes

Comme il est mentionné dans la section précédente, il est important de réaliser que ce que nous faisons référence généralement sous le nom de « Arabe » ne présente pas une seule variété linguistique. Il s'agit d'un recueil de différents dialectes comme l'illustre la figure 1.1.

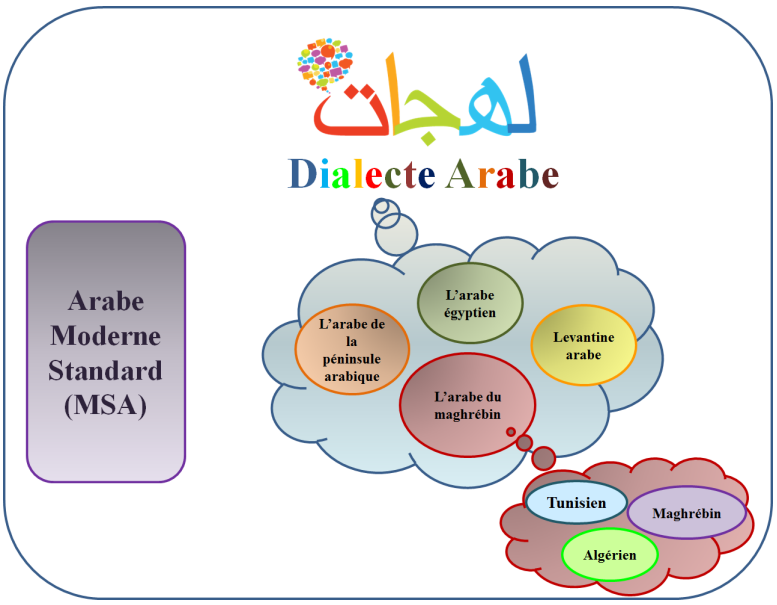

Figure 1.1 – Les variétés linguistiques de la langue arabe.

Considérablement, ces différents dialectes varient les uns des autres et même du MSA. Ainsi, leurs différences affectent tous les niveaux de la langue : la prononciation, la phonologie, le vocabulaire, la morphologie et la syntaxe. À cet égard,

dans le tableau 1.1 nous allons résumer les plus saillantes différences entre le MSA et les langues locales nommées dialectales.

	Dialectes arabe	**MSA**
Grammaire	- Comportent des règles qui sont sous-jacentes. - Peu de dictionnaires et de livres de grammaire.	-Comporte des règles très strictes de grammaire et de syntaxe. - Possède des dictionnaires et des livres de grammaire.
Vocabulaire	- À un vocabulaire populaire qui varie d'une région à une autre. - À un vocabulaire relativement hétérogène contenant des termes relevant de l'arabe classique, des dialectes régionaux, etc.	- À un vocabulaire standard et très riche - À un vocabulaire qui obéit à des règles strictes - À un vocabulaire commun pour tous les arabes et arabophones.
Prononciation	- Elle est variée.	- Elle se prononce uniformément
Statut	- N'a pas d'écriture stricte - Langue maternelle - N'est pas enseigné	- Langue officielle de l'Etat. - Langue d'enseignement et d'administration.
Champs d'utilisation	- Utilisé en famille.	- Toutes les institutes éducatives - Enseignement religieux.

Table 1.1 – Comparaison entre l'arabe dialectal et le MSA.

Dans les parties précédentes, nous avons traité la situation linguistique de l'arabe en abordant son histoire, sa répartition géographique et ses registres. Puis, nous avons observé les grandes divergences existantes entre les différents dialectes en fonction des classifications géographiques. Après avoir traité les distinctions entre les dialectes arabes et le MSA, nous allons mettre l'accent sur le dialecte tunisien puisqu'il représente la langue de notre travail.

1.3 Le dialecte tunisien

De surcroît, au sein du pays tunisien, la situation langagière se définit par la coexistence de l'arabe classique et l'arabe dialectal tunisien. Ainsi, l'arabe représente le statut de la langue officielle du pays qui est réservée aux situations formelles, restreintes et supranationales, mais reste essentiellement écrite et apprise aux écoles primaires et très souvent utilisée pour se communiquer entre tous les pays du monde arabe. En général, comme tous les arabophones, les tunisiens utilisent le dialecte tunisien dans les milieux familiaux et dans leurs vie quotidienne afin d'assurer l'échange au sein de la communauté. Cependant, ce dialecte qui est spécifique pour chaque région, reste une langue parlée et pas écrite contrairement à l'arabe classique. De ce fait, il est clair et net que l'arabisation concerne la langue secondaire du pays et non pas sa langue maternelle, qui semble une tâche difficile à traiter au sein des populations arabophones.

La Tunisie représente un carrefour de diverses civilisations, ainsi sa culture montre une synthèse de différentes influences cumulées au fil de 3000 ans d'histoire et encore sa position géographique stratégique dans le bassin méditerranéen, l'a mené au cœur du mouvement d'expansion des grandes civilisations de la Méditerranée et des principales religions monothéistes. Dans le cadre de notre travail, nous nous intéressons au dialecte tunisien qui revêt des différentes caractéristiques spécifiques et très contrastées par rapport à la langue arabe.

1.3.1 Historique du dialecte tunisien

Située sur la rive sud de la Méditerranée et au nord de l'Afrique, la Tunisie jouit d'une position géographique stratégique qui fait d'elle un point de jonction entre le monde arabe, l'Afrique et l'Europe. Certes, la Tunisie dispose d'une large ouverture sur la Méditerranée qui est entaillée par de nombreux ports naturels et des bras de mer. Cette diversité géographique a donné lieu à de beaux paysages montrant l'originalité de ce territoire. Avec cette situation géographique privilégiée, la Tunisie est devenue par excellence un trait d'union entre l'Europe et l'Afrique, entre l'Orient et l'Occident, et a accueilli sur son sol de grandes civilisations qui ont été attirées par la richesse de cette terre et par l'importance de sa position stratégique au cœur du bassin méditerranéen ce qui a fait d'elle une terre convoitée depuis la plus haute antiquité.

Son histoire plurielle se caractérise par la diversité et la rivalité des puissances successives dés le dernier millénaire avant J.C. Le simple parcours de l'histoire de ce dialecte nous rend compte sur la diversité des peuples qui ont transité par la Tunisie. En fonction de ce qui est affirmé dans l'histoire, les Berbères sont les premiers habitants connus en Tunisie, ils sont les populations autochtones de

l'Afrique du Nord. Par voie de conséquence, le dialecte tunisien est fortement influencé comme tous les dialectes du grand Maghreb par le berbère (l'Amazighe). Après, elle a été le berceau de la brillante civilisation carthaginoise. Avant l'arrivée des arabes, l'Afrique était sous l'occupation romaine du II e au III e siècle. Ce qui a permis la constitution d'un dialecte inspiré des langues romaines en milieu berbère. Par ailleurs, l'Afrique du Nord avait vécu des invasions étrangères, celle des Phéniciens au début du IIe siècle celle des Carthaginois au VIe siècle qui se sont installés dans des comptoirs côtiers. Dans l'Antiquité, l'alternance des civilisations a eu un impact sur l'enrichissement de la langue berbère qui n'a jamais dispersé et elle est encore utilisée par le peuple tunisien dans leur vie quotidienne. Au VIIe siècle, la Tunisie a vécu l'invasion et les conquêtes islamiques venant de la péninsule arabique aussi l'avènement de l'Islam, qui a intégré ce pays au monde musulman. Par conséquence, les berbères ont progressivement adopté la nouvelle foi et la langue arabe propagée, qui est la langue du livre sacré de l'Islam « le Coran ». Ainsi, la langue arabe devient la langue officielle et l'islam devient la religion principale et officielle de la Tunisie. En 1574, l'Afrique du Nord a annexé comme un nouvel état parmi les états de l'Empire ottoman. En effet, tout au long de cette période, la Tunisie n'a pas perdu son arabité mais, elle était influencée par la langue turque ce qui explique l'introduction de quelques mots turcs pour enrichir la langue. Tout ce qui est dit auparavant, illustre la meilleure homogénéité de la Tunisie sur le plan linguistique parmi tous les états du Maghreb. Au bout du XIXe siècle, l'armée française envahit les côtes tunisiennes. Durant la période coloniale, la langue française s'est répandue dans l'administration, dans l'enseignement comme une langue officielle du territoire mais pas comme une langue de communication quotidienne des Tunisiens. Après l'indépendance, le pays Tunisien a commencé à s'arabiser lentement au niveau du secteur d'enseignement, l'administration et la justice, cependant, cette langue française a continué à bénéficier d'un statut privilégié en tant que seconde langue, même si les secteurs mentionnés précédemment restent longtemps bilingues. Pour conclure, la cohabitation de ces différentes langues ou formes dialectales a donné naissance d'un lieu conflictuel mais en même temps complémentaire au niveau de la concurrence entre l'arabe et le français dans plusieurs domaines notamment dans les mass-médias.

Autres faits historiques survenus ont influencé la langue parlée en Tunisie tels que les interactions pacifiques entre les civilisations comme l'italien, l'espagnol et le français, grâce aux échanges commerciaux dû à l'emplacement stratégique de ce pays qui était connu par sa beauté fascinante, sa nature et son climat méditerranéen. Tous ces facteurs lui ont permis de devenir une destination touristique pour tous les touristes à travers le monde. Aussi d'autres faits d'influence s'imposent sur l'immigration légale vers l'Europe ayant pour but éducatif par exemple. Ou encore, l'immigration clandestine à l'Italie en raison du chômage et de la détérioration de la situation économique. Par conséquent, toutes ces influences ont privilégié jusqu'à

maintenant la connaissance des langues européennes renforcées non pas encore à travers le secteur commercial mais aussi par l'intermédiaire de la télévision et du tourisme. En raison de contact de toutes ces civilisations et de ces langues tout au long de plusieurs siècles, nous pouvons relever jusqu'à aujourd'hui les traces de ces langues utilisées dans le discours des Tunisiens dans la vie quotidienne, où nous pouvons remarquer l'adaptation de certains mots à travers l'emprunt lexical qui a donné à ce dialecte une proportion interculturelle et cosmopolite. De cette façon, le dialecte tunisien est le résultat des interactions entre le berbère, l'arabe littéral et de nombreuses autres langues. Cependant, ces influences le rendent intelligible par les arabophones du Moyen-Orient mais plus facilement compris par les arabophones du Maghreb.

Dans ce qui suit, nous allons présenter dans la figure 1.2 l'origine et la signification de quelques mots empruntés et utilisés dans la vie quotidienne des Tunisiens. Il faut noter que toutes ces expressions et ces mots sont utilisés sans avoir subit de modifications phonologiques.

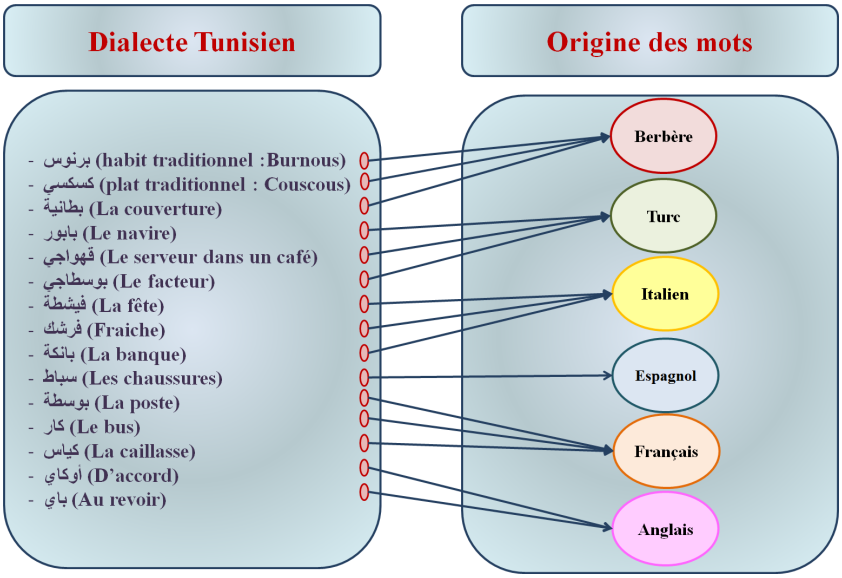

Figure 1.2 – Exemple de mots en dialecte tunisien avec leurs origines et leurs significations

1.3.2 Situation linguistique de dialecte tunisien

D'après nos recherches que nous avons effectuées, nous voulons fournir par la suite la synthèse de la situation linguistique en Tunisie. Pour débuter, le dialecte tunisien représente la langue maternelle utilisée dans la communication courante où s'illustre l'impact du citoyen partagé par tous les Tunisiens quelles que soit

son origine et ses appartenances sociales. Encore, il existe une langue apprise à l'école institutionnellement considérée comme la langue nationale du pays, employée dans les situations formelles, et les Tunisiens, comme la plupart des autres arabophones, l'écrivent mais ne l'utilisent pas oralement, il s'agit de l'arabe littéral. De plus, nous trouvons la langue française qui correspond également à la langue de l'acquisition des sciences et des techniques que nous pouvons découvrir dans certaines universités comme la médecine, des disciplines enseignées entièrement en français. Néanmoins, l'existence de la langue anglaise qui représente la langue internationale, commence à concurrencer le français dans certaines branches de l'enseignement supérieur.

En outre, la langue tunisienne se caractérise par une large ouverture sur d'autres langues étrangères à travers des facteurs historiques et géographiques tels que l'italien et l'espagnol, ou encore par le biais des considérations culturelles allemande et économiques celle des chinois et des russes. À cet effet, cette ouverture se confirme dans l'enseignement supérieur où d'autres langues sont enseignées : l'hébreu, le persan, le japonais, etc.

Pour résumer, historiquement, nous pouvons constater que cette situation linguistique tunisienne évoquée est définie comme étant une situation de diglossie se définit par la coexistence de plusieurs langues à savoir, l'arabe (littéral et dialectal), le français, l'anglais, etc. ce qui approuve l'existence d'un bilinguisme arabe/français dans la société tunisienne auquel s'ajoute une ouverture plus ou moins importante sur d'autres langues.

1.3.3 Registres linguistique de dialecte tunisien

Le paysage linguistique du dialecte tunisien se caractérise par une complexité qui résulte de la rencontre de deux registres linguistique de ce dialecte : le dialecte familier et le dialecte intellectualisé.

- Le dialectal **familier** qui véhicule les besoins quotidiens. En effet, il est utilisé dans la communication orale quotidienne. Nous nous intéressons dans cette thèse à l'étude de ce type de parole.
- Le dialectal **intellectualisé** qui est présent dans les conversations dans les émissions radiophoniques et télévisées [Boukadida 2008]. Ce dialecte se présente comme un mélange entre le MSA et le dialecte tunisien. Ce dernier, quoiqu'il soit énormément stigmatisé et dévalorisé, est bien présent dans les émissions tunisiennes.

1.3.4 Répartition sociolinguistique

À priori, la sociolinguistique est la partie de la linguistique qui étudie les facteurs sociaux déterminant les différences dans la langue et son utilisation par ses

locuteurs. En se basant sur l'observation de ces différences langagières liées à l'âge, au sexe, à la classe sociale..., la sociolinguistique permet de proposer un cadre théorique décrit les différentes variétés qui coexistent au sein d'une communauté linguistique en les mettant en rapport avec les structures sociales (familiales, scolaires, professionnelles, etc.) et explique le fonctionnement social du langage. En effet, comme tous les dialectes, le dialecte tunisien peut être classé selon les différences sociologiques et sociolinguistiques. La diversité du dialecte tunisien d'une région à une autre a créé la distinction entre le ton des villes et le ton de la campagne. De ce fait, nous constatons principalement deux zones dominantes sur le plan linguistique selon la carte géographique de la Tunisie qui sont le Sud "les bédouins" et le Nord. Nous présentons, dans ce qui suit, quelques faits saillants les différentes distinctions du dialecte tunisien entre les régions de Sud et du Nord. Néanmoins, l'existence d'une légère différence au sein de chaque dialecte local au niveau morphologique, phonologique et lexicale n'empêche pas que tous les tunisiens se comprennent les uns les autres, du nord au sud, d'est en ouest.

1.3.4.1 Différences morphologiques

Au niveau morphologique, nous constatons que la distinction de genre n'est pas marquée par toutes les régions Tunisiennes. Citons l'exemple du dialecte du Nord, contrairement au dialecte des autres régions du Sud. Les parlers du Nord ne font pas la distinction entre la deuxième personne du singulier au masculin et au féminin. Le tableau 1.2 montre la conjugaison du verbe قرَأ /qrA/ [Lire] selon ces deux dialectes locaux.

Pronoms	**Dialecte du nord**	**Dialecte du sud**	**Traduction**
« 2MS [7] »	قريت /qryT/	ڤريت /GryT/	Il a lu
« 2FS [8] »	قريت /qryt/	ڤريتي /Gryty/	Elle a lu

Table 1.2 – Exemple de conjugaison du verbe قرَأ /qrA/ [Lire] selon les dialectes du nord et du sud

1.3.4.2 Différences phonologiques

Dans le système consonantique, les régions du Nord utilisent la consonne occlusive sourde ق /q/ [q] alors que les régions du Sud utilisent ڤ /G/ [G]. C'est une propriété qui traduit un clivage sociogéographique entre parler citadin et parler rural [Lajmi 2009].

Dans le système vocalique, le dialecte tunisien se caractérise par l'allongement vocalique. Cette dernière diffère d'une région à une autre, par exemple les parlers des villes du Nord utilisent une voyelle finale longue dans des mots qui portent l'accent sur la dernière syllabe.

Dialecte du nord	**Dialecte du sud**	**Traduction**
سَمَاء /smA/	سَمَاء /sma/	Le ciel
مَاء /mA/	مَاء /ma/	L'eau

Table 1.3 – Exemple de différence du système vocalique entre les régions.

1.3.4.3 Différences lexicales

Le lexique utilisé dans les villes du sud et les autres villes du nord parfois ce n'est pas le même. En effet, nous avons constaté qu'un même référent change de dénomination d'une région à une autre. Le tableau 1.4 illustre ce phénomène.

Dialecte du nord	**Dialecte du sud**	**Traduction**
شتَّاء /$tA/	مطر /mTar/	La pluie
يوقف /ywqif/	يحبس /yaHbis/	S'arrête
منديلة /mindylap/	طبلية /Tabliyap/	Un tablier

Table 1.4 – Exemple de différence au niveau de dénomination entre les régions.

En plus de la variation de dénomination entre les régions Sud et Nord de la Tunisie, nous avons choisi de nous focaliser sur une variété de distinction lexicale qui est l'emprunt. L'emprunt appartient aux moyens dont dispose le locuteur pour enrichir son lexique et pour répondre à un besoin d'expression en utilisant des termes étrangers naturellement intégrés dans le lexique. D'après nos recherches, nous avons remarqué que les Tunisiens de Nord ont utilisé plus les emprunts dans leur discours par rapport à celles de la zone Sud. Cela est expliqué d'une part par l'ouverture des peuples de Nord aux langues étrangères grâce à son emplacement stratégique à proximité des pays européens et d'autre part par l'appartenance de ces peuples à la classe cultivée et instruite.

Dans cette partie, nous avons traité et expliqué les caractéristiques de la répartition sociolinguistique en Tunisie en mettant en évidence ces différenciations. Dans la partie suivante, nous allons exposés les plus saillantes différences entre le MSA et le dialecte tunisien au niveau phonologique, morphologique et syntaxiques.

1.4 Le dialecte tunisien Vs la langue arabe

Comme nous l'avons déjà signalé, le dialecte tunisien représente la langue parlée par les tunisien, c'est leur langage de communication dans les milieux familiaux et d'échange au sein de la communauté. Contrairement aux autres langues à savoir la langue arabe ou la langue Française, ce dialectal ne dispose pas de normes écrites et de descriptions systématiques de ses systèmes phonologique, morphologique et lexical. De ce fait, nous présentons dans les sections qui suivent toutes les caractéristiques du dialecte tunisien en comparaison à l'arabe standard.

1.4.1 Les caractéristiques phonologiques

Le système phonologique de dialecte tunisien est largement altéré par rapport à celui de l'arabe standard. Ces altérations portent en particulier sur les voyelles (suppression ou ajout des suffixes et des préfixes, transformation de leur longueur) [Baccouche 2004] et sur les consonnes. En effet, ce système phonologique se concentre essentiellement sur le système vocalique et consonantique. Alors, nous proposons de présenter les caractéristiques phonologiques en les comparants à l'arabe standard.

1.4.1.1 Système vocalique

Le système vocalique se distingue par : l'allongement vocalique, l'absence des voyelles courtes, changement de voyelle longue. Commençant par l'allongement vocalique, le dialecte tunisien conserve plus ou moins bien le système vocalique de MSA avec ses timbres vocaliques courts et longs [Baccouche 2004]. Néanmoins, il existe certaines variations, à travers l'ajout d'une extension de la durée vocalique, qui peut être apparu dans les mots qui se terminent par la voyelle longue. Généralement, ce phénomène est connu dans les régions nord qu'à celle des régions du sud.

En ce qui concerne l'absence des voyelles courtes, le système phonologique du dialecte tunisien a plusieurs différences par rapport au système phonologique de MSA. En effet, ce dialecte néglige les voyelles courtes surtout quand elles sont situées à la fin d'une syllabe. De plus, la chute de la première voyelle modifie la structure syllabique et exerce une sorte de compactage des unités lexicales qui tendent, pour certaines, vers les monosyllabes [Mejri 2009]. Par exemple, l'exemple du verbe كَتَبَ /kataba/ [il a écrit] en MSA se termine avec la voyelle courte [a]. Par ailleurs, en dialecte tunisien, ce verbe est transformé en كتِب /ktib/ [il a écrit]. Nous constatons l'élimination de la première et la dernière voyelle [a]. Le tableau 1.5 montre la différence entre la prononciation en dialecte et MSA.

Prononciation en dialecte tunisien	**Prononciation en MSA**	**Traduction**
كتِب /ktib/	كَتَبَ /kataba/	Ecrire
صحِن /SHin/	صَحنُ /SahnuN/	Assiette
طَاولة [Tawlap]	طَاوِلَة [Tawilap]	Table

Table 1.5 – Exemple de différence entre la prononciation de quelques mots en dialecte et MSA.

En ce qui touche l'effet de changement de voyelles longues, on remarque l'existence de deux types de changement des voyelles, le premier type consiste à changer la lettre MSA /ay/ en /y/ en dialecte tunisien comme le mot بَيت /bayt/ [maison] se transforme en dialecte tunisien à بيت /byt/. Le deuxième type de changement s'intéresse aux mots contenant à la fin la voyelle longue ى /Y/ qui est généralement raccourcie au niveau de prononciation en une voyelle courte « fatha » comme dans cet exemple : مشَى /m$A/ [marche].

1.4.1.2 Système consonantique

Pour ce qui est du consonantisme, il y a lieu de retenir les faits suivants. L'intervention de la consonne ش /$/ dans le dialecte tunisien partout dans le pays du sud, du centre, du nord, chez les ruraux et chez les citadins. Ceci apparait surtout dans la voix interrogative du dialecte. Les tunisiens disent toujours : وقتَاش /waqtA$/ [quand], علَاش /ElA$/ [pourquoi], كيفَاش /kifA$/ [comment], etc. Ce qui correspond respectivement en MSA à : كيف /kyf/, متَى /mtA/ et لمَاذَا /lmA*A/. Dans le même ordre des idées, le graphème Hamza est multiforme, ses formes sont déterminées par le contexte de la voyelle. En effet, les différentes formes de Hamza qui sont en commun entre MSA et dialecte tunisien sont أ، أي، أُ، آ، أ. Rappelons aussi, l'utilisation de la consonne occlusive sourde ق /q/ dans certaines régions et ڤ /G/ dans d'autres régions.

Après avoir énuméré les différents faits phonologiques du dialecte tunisien, nous passons dans le paragraphe suivant à traiter les caractéristiques morphologiques.

1.4.2 Les caractéristiques morphologiques

Il existe de nombreuses divergences morphologiques entre les dialectes arabes et MSA. Les thèmes dominants sont ceux de la simplification et de l'introduction de nouvelles inflexions clitiques. En ce qui concerne les termes d'inflexions, le cas MSA nominal, l'humeur verbale, le double et le féminin pluriel dans les verbes de conjugaison sont disparus dans le dialecte tunisien. Ainsi, ce dernier introduit de nouveaux clitiques non-MSA que nous allons présenter dans ce qui suit tout en mettant l'accent sur chacune de ces spécificités avec des exemples illustratifs. En fait, la morphologie tunisienne comporte des normes assurant les spécificités de ce langage dialectal. La partie ci-dessous va être décomposée en trois composants le premier est celui de la morphologie verbale, le deuxième traite la morphologie grammaticale et enfin le troisième est consacré pour la présentation de nouveaux clitiques. Traitons la première partie :

1.4.2.1 La morphologie verbale

Ce phénomène marque sa présence notamment dans la différenciation entre le système de la conjugaison des verbes en MSA et celui du dialecte tunisien. Dans ce qui suit nous traitons ces distinctions.

En premier lieu, nous avons constaté qu'il n'existe pas de distinction de genre marquée par toutes les régions de la Tunisie. Citons l'exemple du dialecte du Nord, contrairement au dialecte du sud. Dans cet exemple, la région du nord ne fait pas de distinction entre la deuxième personne du singulier au masculin et au féminin. En effet, nous remarquons la chute du duel masculin et féminin dans la conjugaison du verbe.

En deuxième lieu, le dialecte tunisien a perdu le féminin singulier et le féminin pluriel dans la conjugaison des verbes. Par exemple, passant de MSA au dialecte tunisien, le verbe شَرِبتَ /$aribt/ [tu as bu] en MSA devient en dialecte tunisien شرِبت /$ribt/ qui est le même verbe conjugué en masculin singulier, ainsi nous constatons qu'il n'existe pas de différence entre le féminin et le masculin singulier en dialecte tunisien. De même, il n'y a pas de différence entre le féminin et le masculin pluriel en dialecte tunisien, comme dans l'exemple de verbe conjugué شَرِبتن /$aribtn/[elles ont bu] en MSA devient en dialecte شرِبتوَا /$aribtw/ qui est le même verbe conjugué en masculin pluriel.

En troisième lieu, en dialecte tunisien il n y a pas de différence entre la conjugaison de la première et la deuxième personne du singulier. L'exemple du verbe دخل /dxl/ [entrer] en MSA avec la première personne du singulier أنَا دخلت /AnA

dxlt/ [je suis entré] illustre le même verbe aussi bien ma même prononciation en dialecte tunisien de la première et la deuxième personne du singulier أنَا دخلت /AnA dxlt/ [je suis entré] et أنت دخلت /Ant dxlt/ [tu as entré].

Enfin, la forme passive du verbe : À priori, la passivation en dialecte tunisien des verbes trilitères est produite à travers la précession de la consonne ت [t] dans le verbe à l'accompli. Contrairement au MSA, la passivation des verbes engendre la transformation de la structure de la phrase comme l'illustre cet exemple. La phrase en MSA dans la forme passive : أكِلَت التِفَاحَة /Akilat AltfAHap/ [la pomme est mangée]. Cependant, la phrase en dialecte tunisien devient dans la forme passive : التِفَاحَة تَاكِلت /AltfAHap tAklt/ [La pomme a été mangée].

En revanche, dans l'inaccompli, le ت [t] se trouve entre la racine et le marqueur préfixe tendu selon le genre masculin ou féminin comme dans ce qui suit : يتَاكل /ytAkil/ [manger] en masculin et تِتَاكل /tetAkil/ en féminin. Néanmoins, en dehors des formes passives, le dialecte tunisien propose une autre forme plus souvent utilisée dans la communication quotidienne qui est la forme interrogative comme par exemple تِتَاكلشي /titAkil$y/ .

1.4.2.2 Catégorie grammaticale

Dans cette phase nous pouvons distinguer deux types de catégories grammaticales : catégorie du nombre et catégorie du genre. Pour la catégorie du nombre, il faut indiquer que le duel a presque disparu du dialectal au niveau du numéral زوز /zwz/ [deux] préfixé ou suffixé [Mejri 2009]. Par exemple, les deux formes préfixé et suffixé du mot زوز sont زوز أقلَام /zwz AqlAm/ ou أقلَام زوز /AqlAm zwz/ [deux stylos]. Cependant, en MSA le duel prend cette forme « le nom au singulier + le suffixe ي /y/» comme dans cet exemple, قلمين /qalamayn/ [Deux stylos].

Quant au catégorie du genre comporte des formes ambigües comme le mot en MSA عروس /Erws/ [mariée] devient en dialecte sera عروسَة /Erwsap/. De plus, il n'existe pas de distinction entre le féminin et le masculin. Par exemple, le mot أرنب /Arnb/ [Lapin] désigne le féminin et le masculin en même temps.

1.4.2.3 Les nouveaux clitiques

Le dialecte tunisien introduit de nouveaux clitiques inexistants en MSA tel que le clitique de la négation مَا + ش /mA $/ [ne pas] qui se manifeste en MSA avec de différentes particules à savoir, مَا /mA/, لَا /lA/, لَن /lan/ et لَم /lam/ [pas]. Généralement, ces particules en MSA se situent souvent devant le verbe et peuvent parfois modifier la conjugaison comme le verbe مشَى /m$A/ [aller] lorsqu'il est précédé d'une particule de négation tel que لم /lam/ [ne pas], il donne naissance à l'expression de la négation لم أمشي /lam Am$y/ [je n'irai pas]. Néanmoins, en dialecte tunisien, ce verbe change en مَا مشيتش /mA myti/ en lui ajoutant la particule de négation مَا /mA/. Dans un autre cas de le MSA, on trouve le clitique de l'interrogation verbale أ /A/ et la particule هل /hal/ qui sont remplacés par le clitique شي /$y/ dans le dialecte tunisien.

Après avoir exposé quelques faits saillants de la morphologie de ce dialectal, nous allons étudier toutes les caractéristiques lexicales du dialecte tunisien dans la section qui suit.

1.4.3 Les caractéristiques lexicales

Le système lexical du dialecte tunisien semble un système plus ouvert ayant un vocabulaire spécifique très riche. En effet, nous avons choisi de nous focaliser sur un certain nombre de phénomènes systémiques tels que l'emprunt et la dérivation.

L'emprunt, ceci est un autre domaine où le dialecte tunisien illustre un excellent dynamisme. Nous distinguons les trois points suivants. L'intégration de nouveaux suffixes empruntés d'autres langues tel que le suffixe «-iste » d'origine française dont certains emplois dépassent les emprunts pour s'appliquer à des bases d'origine arabe [Mejri 2009]. Par exemple : قرَاجيست [garajisto] /quelqu'un a un garage/. Aussi, l'impact phonologique qui agit par le biais de l'emprunt sur le système phonologique du dialectal. L'aspect qualitatif se mesure par ailleurs au moyen de l'impact que les emprunts pourraient avoir sur le système phonologique du dialectal [Mejri 2009]. Nous fournissons ici deux exemples où une seule consonne sera changée : ميكَانيسيَان /mykAnysyAn/ [mécanicien] et الكتريسيَان /lAktrysyAn/ [électricien].

De même, l'intégration systématique des unités empruntées dans les paradigmes construits par schèmes. Autrement dit, à partir d'un mot emprunté, nous pouvons avoir facilement toutes les unités (verbe, nom, agent, etc.) répondant à tous les schèmes disponibles en dialectal [Mejri 2009]. Dans le tableau suivant nous allons montrer cette intégration :

Le mot emprunté	Le verbe	Le nom	L'agent
[bɛznɛs] « bisness » « بزنس»	[bɛznɛsɛ] « il a fait un biseness » « بيزنس»	[tbɛezni:s] « action de faire des biseness » « تبزنيس»	[bɛznɛ:s] (singulier) « بزناس » [beznɛ:sa] (pluriel) « بزناسا» « celui qui fait du biseness »

Figure 1.3 – L'ensemble des unités du mot d'origine française « bisness ».

La dérivation de l'arabe littéral est d'une grande régularité. Ainsi, à partir d'une racine consonantique, on peut dériver selon des schèmes préétablis, impliquant une variation vocalique et l'ajout de certains éléments consonantiques, un ensemble de paradigmes exprimant l'agent, le patient, le locatif, les noms prédicatifs, le superlatif, etc. [Mejri 2009]. De même pour le dialecte tunisien, plusieurs paradigmes d'unités suffixées connaissent un enrichissement continu, notamment grâce à l'incorporation dérivationnelle [Sfar 2005] et un certain nombre d'affixes spécifiques comme جي /jy/ qui indique la profession [Baccouche 1994] : قهوَاجي /qahwAjy/ [celui qui possède une cafétéria], بنكَاجي /bankAjy/ [banquier] et كوَارجي /kwArjy/ [footballeur].

Pour conclure l'énumération des caractéristiques du dialecte tunisien, nous allons présenter dans la partie suivante les caractéristiques syntaxiques de ce dialecte.

1.4.4 Les caractéristiques syntaxiques

Il est nécessaire de noter que la syntaxe représente le domaine où la rupture avec le MSA est la plus importante. Dans cette rubrique, les outils syntaxiques dialectaux connaissent deux tendances contradictoires [Mejri 2009] que nous allons présenter par la suite.

Commençons par la première tendance, cette dernière consiste à réduire soit le nombre soit le forme des pronoms de MSA [Mejri 2009] de douze à sept pronoms personnels. Aussi, nous avons rendu compte de la disparition du duel de la deuxième personne أنتمَا /AntmA/ [vous les deux] et celui de la troisième personne

همَا /hmA/ [ils], le pluriel féminin de la deuxième personne أنتنّ /Antn/ [vous] et le pluriel féminin de la troisième personne هنّ /hn/ [elles]. De même pour les pronoms démonstratifs qui ont perdu la forme duelle au féminin هَاتين /hAtyn/ et au masculin هذين /h*yn/ [Ces]. Par contre les pronoms relatifs ont gardé leur forme en dialecte tunisien quelque soit leur genre qui est la forme unique اِلّي /Qui/.

Passons à la deuxième tendance, elle se caractérise par une tendance inverse par rapport à la première dont les pronoms démonstratifs illustrent la pluralité [Mejri 2009]. Comme par exemple, en MSA : هُولَاء [ceux] devient en dialecte tunisien هَاذم [hA*m] et هَاذومَ.

1.5 Conclusion

Dans ce chapitre, nous nous sommes intéressés à présenter les registres linguistiques de la langue arabe ainsi que le passage historique de l'arabe classique vers le MSA jusqu'à l'arrivée à l'arabe dialectal. De plus, nous avons donné quelques traits historiques de l'apparition du dialecte tunisien et définir sa situation linguistique. Enfin, nous avons clôturé ce chapitre par citer les plus saillantes différences entre le MSA et le dialecte tunisien à ses différentes niveaux : le niveau morphologique, le niveau phonologique et le niveau lexical.

Compte tenu de la complexité linguistique du dialecte tunisien décrite ci-dessous et qui se définit en tant qu'une « *diglossie* » [Fishman 1967] où plusieurs langues et leurs variétés coexistent, ce dialecte se plaint d'un manque de ressources et d'outils nécessaires pour son traitement. De ce fait, ce défi constitue notre principale motivation pour la conception et la réalisation d'un SRAP pour le dialecte tunisien. Ainsi, la description de ce travail fait l'objet de la suite de cette thèse.

Dans le chapitre 2, nous allons présenter un aperçu sur les différents composants d'un SRAP allant de l'acquisition informatique du signal de parole jusqu'à sa transcription.

Chapitre 2

Reconnaissance automatique de la parole

Sommaire

2.1 Introduction

Dans notre perspective de développement d'un système de reconnaissance automatique de la parole (SRAP) en dialecte tunisien, il est essentiel d'avoir une bonne vision d'ensemble du processus permettant de passer d'un signal audio à sa transcription textuelle. L'objectif de ce chapitre est donc de dresser un panorama tant théorique que pratique du fonctionnement d'un SRAP.

Ce chapitre est réparti en deux principales parties. La première partie est composée de huit sections. Après la présentation du principe général d'un SRAP, nous consacrons la section deux à la description de l'étape de paramétrisation du signal de la parole. La troisième section illustre le principe de fonctionnement du module acoustique fondé sur l'utilisation des modèles de Markov cachés, notés HMM pour Hidden Markov Models. Dans la quatrième section, nous présentons, en premier lieu, les modèles de langage les plus connus et en deuxième lieu, nous exposons la mesure d'évaluation la plus utilisée pour estimer les performances des modèles de langage. Le dictionnaire de prononciation fera l'objet d'une cinquième section. Nous conclurons cette première partie par une description de métrique d'évaluation des SRAP.

Dans la deuxième partie de ce chapitre, nous présentons un aperçu sur quelques SRAP développés dans la littérature pour les langues peu dotées, allant de l'acquisition des ressources textuelles et de parole jusqu'aux expérimentations.

I. Architecture d'un système de reconnaissance automatique de la parole

2.2 Principes généraux

Un Système de Reconnaissance Automatique de la Parole (SRAP) a pour objectif la transcription textuelle d'un signal de la parole. En effet, dans le cadre de la

modélisation statistique de la parole, cette tâche est l'équivalent de construire des séquences de mots possibles W* étant donnée une séquence X de caractéristiques acoustiques observées à partir du signal d'entrée. Mathématiquement, cela s'écrit sous la forme suivante :

$$W^* = arg_W max\, P(W|X) \tag{2.1}$$

La probabilité P(W|X) est très difficile à déterminer [Ueberla 1994], d'où la nécessité de la décomposer. En utilisant la règle de Bayes, il est possible de reformuler la probabilité (W|X) comme suit :

$$W^* = arg_W\, max \frac{P(W|X)P(W)}{P(X)} \tag{2.2}$$

Où P(X|W) est la vraisemblance des observations acoustiques sachant une séquence de mots testée, P(W) est la probabilité a priori de cette séquence de mots et P(X) est la vraisemblance a priori de la réalisation acoustique. Puisque la vraisemblance P(X) est considérée constante quelque soit la séquence W, elle est donc retirée de l'équation 3.2. La recherche de W peut se simplifier en :

$$W^* = arg_W\, max\, P(W|X)P(W) \tag{2.3}$$

D'après cette équation, le point visé de chaque SRAP peut être résumé comme suit : la recherche de la séquence de mots la plus probable W. Pour ce faire, deux types de modèles probabilistes sont utilisés : un modèle de langage qui fournit la valeur de P(W) et un modèle acoustique qui fournit la valeur de P(X|W). La première valeur représente la probabilité a priori d'observer la suite de mots W indépendamment du signal. La deuxième valeur indique la probabilité d'observer la séquence de vecteurs acoustiques X sachant une séquence de mots spécifiques W.

Schématiquement ces étapes peuvent être résumées d'une manière simplifiée par la figure 2.1.

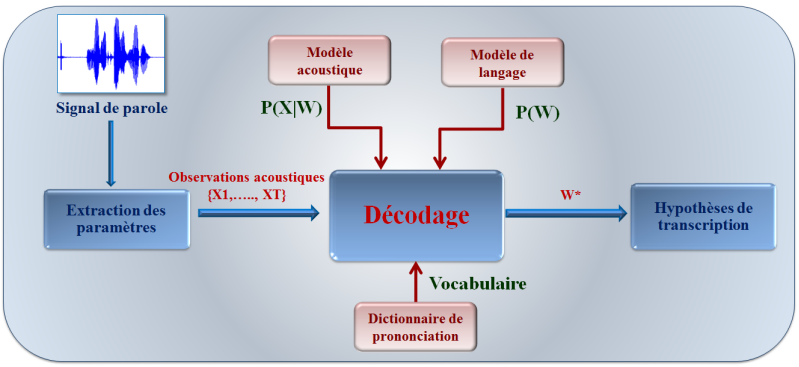

Figure 2.1 – Vue schématique d'un Système de Reconnaissance Automatique de la Parole.

La figure 2.1 montre les différentes étapes nécessaires à la reconnaissance d'un message m prononcé en entrée. En premier lieu, le signal de parole est subdivisé en vecteurs acoustiques. En deuxième lieu, et en utilisant ces vecteurs, le modèle acoustique se charge de construire la suite des phonèmes hypothèses du signal prononcé. En troisième lieu, la suite de mots obtenue sera évaluée par le modèle de langage qui permet d'estimer la probabilité P(W). Ce processus est répété pour toutes les hypothèses possibles. En dernier lieu, le système donne les$N$ meilleures hypothèses comme résultat de la reconnaissance.

Les sections suivantes de ce chapitre sont dédiées à la description des trois grandes étapes de reconnaissance qu'on vient de mentionner.

2.3 Extraction de paramètres

La variabilité et la redondance du signal de la parole le rendent difficilement exploitable, tel qu'il est dans un SRAP. Pour contourner ce problème, il est nécessaire d'en extraire uniquement les paramètres qui seront dépendants du message linguistique et utiles pour la reconnaissance. Pour ce faire, le signal est tout d'abord découpé en trames. Chaque trame est considérée comme une fenêtre de 10 à 20 ms de signal. Dans cet intervalle, on suppose que le signal vocal est suffisamment stable. Par la suite, un vecteur de paramètres acoustiques est extrait pour chacune de ces trames. Suite à ces prétraitement, on obtient une séquence d'observations acoustiques X, où $\mathbf{X} = \mathbf{x_1 x_2 ... x_m}$, et chaque vecteur $\mathbf{x_i}$ représente quelques millisecondes (typiquement 10 ms).

Dans la littérature, il existe plusieurs méthodes de paramétrisation pour extraire uniquement les paramètres qui seront dépendants du message linguistique. Les méthodes les plus couramment utilisées de nos jours font appel à l'analyse

cepstrale, telles que la méthode Mel-scale Frequency Cepstral Coefficients (MFCC) [Davis 1990] ou la méthode Perceptual Linear Prediction (PLP) [Hermansky 1991] et les LPCC Linear Prediction Cepstral Coefficients (domaine temporel) [Markel 1982]. Le jeu de paramètres obtenu est couramment augmenté par leurs dérivées premières ($\boldsymbol{\Delta}$) et secondes ($\boldsymbol{\Delta\Delta}$) qui permettent de mieux modéliser les caractéristiques dynamiques des paramètres acoustiques (vitesse et accélération).

2.4 Modélisation acoustique

Le rôle d'un modèle acoustique est de calculer la vraisemblance P(X|W) définie dans l'équation. Les modèles de Markov cachés (Hidden Markov Model - HMM) sont les plus communément utilisés en modélisation acoustique pour estimer cette probabilité. Ces HMM sont des automates stochastiques à états finis capable de calculer la probabilité d'émission d'une séquence d'observation donnée.

L'objectif des HMM est donc de modéliser au mieux les unités représentatives du signal de la parole. Les unités les plus utilisées sont les phonèmes. Néanmoins, il existe aussi d'autres types d'unités représentatives du signal de la parole à savoir, les allophones, les syllables, les triphones, les mots, etc.

Dans la suite, nous allons définir mathématiquement et schématiquement les modèles de Markov cachés. Après, nous expliquons comment mettre en pratique la théorie des HMM. Finalement, nous mettons l'accent sur les limites des HMM.

2.4.1 Définitions des modèles de Markov cachés

Un HMM peut être vu comme un ensemble discret d'états et de transitions entre ces états. Formellement, il peut être défini par l'ensemble des paramètres [Rabiner 1989] :

$$\lambda = (S, \Sigma, T, G, \pi) \tag{2.4}$$

Où :

- **S** est un ensemble de $\boldsymbol{N}$ nœuds ou d'états.
- $\sum$ est un alphabet de $\boldsymbol{M}$ symboles.
- **T** est la matrice des probabilités a_{ij} de transition entre les états, de taille $N \times N$. La somme des probabilités de transitions entre un état $\boldsymbol{i}$ et tous les autres états doit être égale à **1**, *i.e* $\forall i, \sum_{j=1}^{N} a_{ij} = 1$.

- $G = S \times \Sigma \rightarrow [0, 1]$ est la matrice d'observation indiquant les probabilités de génération associées aux états, $b_i(o_t)$ est la probabilité de générer le symbole $o_t \in \Sigma$ à partir de l'état $\boldsymbol{i}$. La somme des probabilités des émissions partant d'un état est égale à $\mathbf{1}$, $i.e\ \forall i, \sum_{o_t} bi(o_t) = 1$.
- $\pi : S \rightarrow [0, 1]$ est un vecteur de probabilités initiales. La somme de ces $\boldsymbol{N}$ probabilités doit être égale à $\mathbf{1}$, $i.e\ \forall i, \sum_i \pi_i = 1$.

Schématiquement un HMM peut être représenté comme suit :

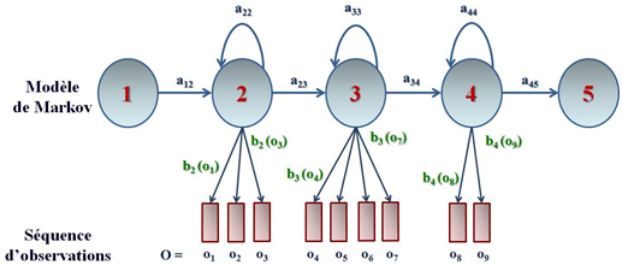

FIGURE 2.2 – HMM à 5 états dont 3 émetteurs.

La figure 2.2 illustre un exemple d'un HMM à 5 états dont 3 états émetteurs utilisés pour la modélisation d'un phonème et 2 états utilisés pour désigner les états d'entrée et de sortie. Ces deux derniers sont ajoutés dans le but de faciliter la concaténation des modèles entre eux. En effet, l'état de sortie d'un modèle de phonème peut être fusionné avec l'état d'entrée d'un autre modèle de Markov caché pour former un modèle composite. Partant de ce fait, les modèles de phonèmes peuvent être concaténés ensemble pour former les mots et ainsi les phrases.

Un HMM est considéré comme un générateur de vecteurs acoustiques. Un HMM peut être considéré comme une machine à états finis qui change d'état à chaque unité de temps. Pour chaque unité de temps $\boldsymbol{t}$, une fois arrivé à l'état q_j, un vecteur acoustique (o_t) est généré avec une densité de probabilité $b_j(o_t)$. De plus, la transition de l'état q_i à l'état q_j est probabiliste, sa probabilité est généralement notée a_{ij}. En pratique, c'est seulement la séquence d'observations $O = o_1, o_2 ... o_T$ qui est connue.

2.4.2 Les limites des HMM

Etant donné un modèle de Markov caché λ et une séquence d'observations acoustiques O, la reconnaissance de cette séquence s'effectue en trouvant le modèle λ qui maximise la probabilité P $(\lambda|O)$ (probabilité qu'un modèle λ génère une séquence d'observations O. Cependant, il n'est pas possible d'accéder directement à cette probabilité, mais on peut calculer la probabilité qu'un modèle donné générera une certaine séquence d'observations P$(O|\lambda)$.

En utilisant la loi de Bayes, il est possible de lier ces deux probabilités par :

$$P(\lambda|O) = \frac{P(O|\lambda)P(\lambda)}{P(O)} \tag{2.5}$$

Où :

- $P(O|\lambda)$ est la vraisemblance de la séquence d'observations **O** étant donné le modèle λ.
- $P(\lambda)$ est la probabilité a priori du modèle.
- $P(O)$ est la probabilité a priori de la séquence d'observations.

La probabilité $P(O)$ peut être considérée constante, sous prétexte qu'elle est indépendante du modèle λ, si les paramètres de ce dernier sont fixés. En effet, maximiser $P(\lambda|O)$ revient à maximiser $P(O|\lambda)\ P(\lambda)$.

Les trois problèmes suivants sont soulèves par l'utilisation des HMM, qui vont être abordés dans les parties suivantes :

- Le premier problème est un problème d'évaluation, qui peut également être vu comme un problème d'estimation de la capacité d'un modèle donné à reconnaitre une séquence d'observations donnée. Autrement dit, étant donné un modèle de HMM $\lambda = (S, \Sigma, T, G, \pi)$, quelle est la probabilité $P(O|\lambda)$ d'avoir une séquence d'observations $O = o_1, o_2...o_T$?
  Ce problème a été remédié par l'algorithme Forward-Backward. Les détails de cet algorithme sont exposés dans [Juang 1992].
- Le second problème, est un problème de décodage, se ramène à l'idée de dévoiler les états cachés S, sans $\boldsymbol{y}$ avoir accès directement. Dans la plupart des cas, le critère d'optimalité retenu influencera la séquence d'états calculée. Formellement, le problème est se décrit comme suit :
  Étant donnée une séquence d'observations $O = o_1, o_2...o_T$ et le modèle $\lambda = (S, \Sigma, T, G, \pi)$ comment trouver la séquence d'états qui maximise la séquence d'observations ?
  Ce problème a été résolu par le recours à l'utilisation de l'algorithme de Viterbi. Les détails de cet algorithme sont exposés dans [Juang 1992].
- Enfin, le troisième problème se ramène à l'entraînement d'un HMM par des séquences d'observations, en vue d'en optimiser les paramètres pour un problème spécifique donné. Ceci, nous ramène à la question suivante : comment déterminer les paramètres du modèle $\lambda = (S, \Sigma, T, G, \pi)$ afin de maximiser $P(O|\lambda)$?
  Ce problème a été remédié par l'algorithme EM (Expectation-Maximization). Pour une présentation détaillée de cette méthode d'apprentissage, nous invitons le lecteur à consulter [Juang 1992].

2.5 Modélisation statistique de langage

Les modèles de langage visent à représenter le comportement de la langue afin de confirmer ou infirmer les propositions faites par le module acoustique. On distingue deux grandes familles de modèles de langage, les modèles à base de grammaires et les modèles probabilistes. Ces derniers sont ceux qui dominent actuellement le domaine de la reconnaissance de la parole. Un modèle de langage probabiliste est construit à partir d'un grand corpus d'apprentissage composé de données exprimées dans la langue étudiée. Il a pour but d'estimer la probabilité $P(W)$ où W est une suite de mots $W = w_1, w_2, ..., w_N$.

Dans la littérature plusieurs approches statistiques pour la modélisation du langage sont reconnues comme étant les plus performantes en reconnaissance automatique de la parole. Ces approches sont fondées sur l'estimation de probabilités de n-grammes, séquences de n mots. Nous allons exposer chacune de ces approches dans les sous-sections suivantes.

2.5.1 Le modèle n-grammes

Grâce à leur simplicité et à leur efficacité, les modèles n-grammes constituent les modèles de langage les plus employés dans le domaine de la reconnaissance de la parole. Ils sont basés sur l'hypothèse que l'apparition d'un mot dépend seulement de son historique. Détaillons donc cette méthode de modélisation du langage. Le modèle de langage n-grammes permet d'estimer la probabilité a priori $P(W)$ d'une séquence de mots. Formellement, la probabilité de la séquence de mots $w_1, ..., w_i, ..., w_N$ est calculée par :

$$P(w_1, w_2, ...w_N) = P(w_1) \times \prod_{i=2}^{N} P(w_i|w_1...w_{i-1}) \tag{2.6}$$

avec $P(w_1)$ la probabilité d'observer le mot w_1 et $P(w_i|w_1...w_{i-1})$ celle de rencontrer le mot w_i après la séquence $w_1, ...w_{i-1}$. En pratique, l'estimation de cette probabilité est très difficile. De fait que, aucun corpus d'apprentissage ne peut permettre d'observer toutes les suites de mots possibles. De ce fait, l'idée de base des modèles n-grammes consiste donc à ne considérer que les suites de mots de longueur n c'est à dire le calcul est approché par un historique limité constitué des n-1 mots précédents. Le calcul de ces probabilités est basé sur le comptage de chaque séquence observée. L'ordre de modèle de langage définit la taille de l'historique n-1 à prendre en considération dans le calcul. Ainsi, dans le cadre d'un système de reconnaissance automatique de la parole à grand vocabulaire, $\boldsymbol{n}$, l'ordre du modèle, est typiquement compris entre 2 et 5.

L'inconvénient majeur de ce type de modélisation est que l'application de la formule 2.6 conduit à attribuer une probabilité nulle à tout n-gramme n'ayant jamais été rencontrée dans le corpus d'apprentissage. Ce problème est spécialement grave quand ce n-gramme pourrait être parfaitement valide sur le plan linguistique. Il devient donc nécessaire d'utiliser d'autres techniques pour estimer les probabilités des séquences de mots non rencontrées. Quelques une de ces techniques seront présentées dans la sous-section suivante.

2.5.2 Techniques de lissage

Aussi long et représentatif qu'il soit, un corpus d'apprentissage ne peut contenir toutes les suites possibles des mots. De sorte que, plusieurs modèles de langage attribuent une probabilité nulle à tout n-gramme n'ayant jamais été rencontré dans le corpus d'apprentissage.

Les techniques de lissage tentent de compenser cette carence. Ces techniques consistent à attribuer une probabilité non nulle aux mots et séquences de mots non rencontrés. Le principe général de ces techniques consistent à prélever une quantité à la masse des probabilités issue des évènements observés et de la redistribuer aux probabilités associées aux évènements non ou peu vus. Dans la littérature, souvent de méthodes de lissage sont utilisées pour l'estimation des données manquantes et pour le lissage des probabilités des modèles de langage. Les plus notables de ces méthodes, sont le lissage de Good-Turing [Good 1953], de Witten-Bell [Witten 1991] et de Kneser-Ney [Kneser 1995].

2.5.3 Modèles de langage n-classes

En raison du manque de données d'apprentissage, il est nécessaire de trouver une méthode qui permet de maximiser la quantité d'information utile d'une part et de réduire l'espace de paramètres du modèle d'autre part.

Afin de répondre à cette exigence, d'autres méthodes ont vu le jour. Elles consistent à regrouper les mots en classes. Ceci correspond à l'apparition d'une modèle de langue de type n-classes (class-based language models). L'idée principale de ce modèle est de regrouper les mots du vocabulaire en classes lexicales et de considérer principalement le calcul de la probabilité d'une séquence de mots comme celui de la probabilité d'une séquence de classes lexicales [Brown 1992].

En effet, les modèles n-classes consistent à attribuer à chaque mot w_i une classe $C(w_i)$ et à estimer les probabilités des mots en fonction de deux facteurs : la probabilité d'appartenance du mot à sa classe $P(w_i|C(w_i))$ et la probabilité d'apparition de cette classe à la suite de son historique de classes. Dans le cas où un mot ne peut appartenir qu'à une seule classe, la probabilité du mot w_i sachant son historique $w_{i-n+1}, ...w_{i-1}$ est définie comme suit :

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

Le terme $P(C(w_i)|C(w_{i-n+1}), ..., C(w_{i-1}))$ qui correspond à la probabilité de la succession des classes $C(w_{i-n+1}), ..., C(w_{i-1}), C(w_i)$ est estimé de la même manière que la probabilité $P(w_i|w_{i-n+1}, ..., w_{i-1})$. La probabilité d'appartenance du mot W_i à la classe $C(W_i)$, quant à elle, est calculée selon la formule suivante :

$$P(w_i|C(w_i)) = \frac{N(w_i)}{N(C(w_i))} \tag{2.7}$$

où $N(...)$ est le nombre d'occurrences de l'argument dans le corpus et $C(...)$ est la classe à laquelle appartient le mot en argument.

Plusieurs variantes de modèles de type n-classes sont présentes dans la littérature. Nous pouvons les décomposer en quatre catégories : les classes syntaxiques qui regroupent les mots selon leur catégorie grammaticale, les classes morphologiques qui regroupent les mots ayant la même racine morphologique (lemme), les classes sémantiques qui regroupent les mots ayant le même sémantique et enfin les classes obtenues par d'autres méthodes de classification automatique.

Une des motivations les plus claires pour les modèles n-classes est le fait qu'un mot d'une classe donnée, ne se trouvant pas forcément dans le corpus d'apprentissage, hérite de la probabilité de tous les autres représentants de sa classe. De plus, il est possible, d'ajouter des mots dans les classes, sans avoir besoin de ré-estimer les probabilités du modèle.

Cependant, les problèmes auxquels sont confrontés les modèles de type n-classes sont nombreux. Dans ce qui suit nous citons, de façon non exhaustive, quelques problèmes les plus rencontrés au sein de ce type de modèles. La première grande difficulté est que ce type de modèle exige la nécessité d'avoir un corpus d'apprentissage pré-étiqueté [Smaïli 1991]. Néanmoins, l'étiquetage manuel, malgré qu'il donne des résultats exacts, il est particulièrement lourd. La deuxième difficulté, partagée entre les modèles n-classes et les modèles n-grammes, concerne la taille de l'historique pris en compte. En fait, la majorité des modèles présentés dans la littérature se limitent à un historique restreint à 3 ou 4 classes, ce qui est insuffisant pour une modélisation fiable du langage.

2.5.4 Autres modèles de langage

Dans les sections précédentes, nous avons détaillé quelques types de modèles de langages qui nous semblent les plus utilisés dans la littérature. Cependant, il ne faut pas perdre de vue que bien d'autres approches sont proposées. Notamment, certains travaux, dans la famille des modèles n-grammes, suggèrent d'utiliser des historiques de longueurs variables ([Niesler 1996] ;[Bonafonte 1996]), de sauter ou d'inverser certains mots de l'historique ([Rosenfeld 1994] ; [Langlois 2000]) ou de mélanger ces différentes modélisations [Brun 2007].

Après avoir exposé quelques types de modèle de langage dans la littérature, la prochaine section aborde la question d'évaluation de ces modèles.

2.5.5 Evaluation d'un modèle de langage

Indépendamment de leurs différences et de leurs variétés, l'intention principale d'un modèle de langage est de construire des distributions de probabilités conditionnelles sur les mots d'un vocabulaire. Ceci permet de poser un cadre d'évaluation commun qui permet de comparer la qualité de la modélisation linguistique de deux modèles de langue. Cependant, évaluer un modèle de langue statistique sur les sorties d'un système de reconnaissance automatique de la parole ne cible pas l'évaluation de la seule qualité de la modélisation linguistique étant donné que le processus de transcription dépend de nombreux autres facteurs, notamment de la nature acoustique. De plus, la nécessité de lancer un SRAP complet pour évaluer un modèle de langage s'avère très coûteuse contraignante.

En effet, une mesure alternative simple et efficace a été introduite afin d'évaluer les modèles de langage sans les intégrer dans un système de reconnaissance. Il s'agit de la mesure de perplexité (PPL). Cette mesure est un indicateur de la capacité de prédiction du modèle de langage. Son principe est de vérifier à quel point un modèle de langage est capable de prédire les séquences de mots du langage qu'il est censé modéliser. En pratique, la PPL calcule, pour chaque position d'un mot W, le nombre moyen de choix possibles [Bougares 2012]. La perplexité est calculée au moyen de la formule suivante :

$$PPL = 2^{-\frac{1}{n}\sum_{t=1}^{n}\log_2 P(w_t|h)} \tag{2.8}$$

Où $P(w_t|h)$ représente la probabilité proposée par le modèle de langage pour le mot w_t sachant l'historique h.

Une perplexité de valeur K est interprétée comme suit : plus la valeur de $\boldsymbol{K}$ obtenue est faible (le but étant de se rapprocher de 0) plus le modèle de langage possède de meilleurs capacités de prédiction.

Grâce à son simplicité de calcul, la perplexité est une mesure très communément utilisée pour évaluer un modèle de langage. Autrement dit, cette mesure de perplexité donne une appréciation moyenne de la qualité d'un modèle de langage. Toutefois, les expériences montrent que la perplexité ne doit pas être vue comme une valeur corrélée au taux de reconnaissance d'un système. En effet, il arrive souvent qu'un modèle L' ayant une perplexité meilleure qu'un autre $\boldsymbol{L}$ ne permette pas, une fois intégré dans un système de reconnaissance, d'améliorer les performances obtenues initialement par $\boldsymbol{L}$. Aussi, dans la littérature il y a des travaux [Clarkson 1999] qui montrent qu'une diminution de la perplexité ne se traduit pas par une diminution du **WER**.

2.6 Dictionnaire de prononciation

Le lien entre la modélisation acoustique et la modélisation linguistique est fait par un dictionnaire de prononciation peut être appelé aussi dictionnaire de prononciation. Un tel dictionnaire doit contenir un vocabulaire qui peut être défini comme l'ensemble des mots qu'un SRAP est capable de reconnaître, ainsi que leurs prononciations.

Autrement dit, il est nécessaire d'associer chaque entrée du dictionnaire à une suite de phonèmes qui lui est propre. Un phonème peut correspondre à un ou plusieurs lettres (graphèmes) différents, cela signifie qu'il est nécessaire de disposer de toutes les séquences de phonèmes correspondant à un mot dans le dictionnaire. Pour convertir les graphèmes en symboles phonétiques, un système appelé conversion graphème en phonème (G2P) est utilisé. Dans la littérature, il existe trois approches qui ont été utilisées pour la transcription phonétique d'un mot donné à savoir, l'approche manuelle, l'approche à base de règles où la connaissance linguistique et phonétique des experts est utilisée pour développer un ensemble de règles et l'approche guidée par les données. Plus de détails sur les approches de conversion G2P sont donnés au chapitre 3.

2.7 Décodeur

La dernière étape dans le processus de reconnaissance automatique de la parole consiste à intégrer les résultats de la modélisation acoustique et ceux des modèles de langage dans un seul processus de décision permettant de retrouver un message prononcé en entrée. Pour ce faire, un composant principal d'un SRAP est mis en place, il s'agit d'un "décodeur".

Partant des informations contenues dans le dictionnaire de prononciation et des modèles acoustiques et linguistiques, la question de décodage d'un signal de la parole consiste à parcourir l'espace de recherche que représente l'intégralité des

séquences de mots possibles à partir du vocabulaire du système et à trouver le meilleur chemin qui donnera la séquence de mots la plus probable. Clairement, cet espace de recherche, est très grand, et représenté sous forme de graphe appelé "graphe de recherche". En plus, il intègre certaines informations utilisées pour générer les hypothèses telles que les unités acoustiques (phonèmes) associées à leurs scores acoustiques.

Etant donné que l'espace de recherche est très grand et dans le but de n'explorer qu'un espace restreint et suffisant pour trouver la meilleure solution, des algorithmes de recherches sont employés pour choisir à chaque instant un nombre limité d'hypothèses. La stratégie de recherche, dans cet espace d'hypothèse, diffère d'un algorithme à un autre. En effet, il y a des algorithmes qui incluent la stratégie de recherche en "profondeur d'abord" alors que d'autres incluent la stratégie en " largeur d'abord". Ces deux types de stratégies sont relatifs à des parcours différents d'une arborescence de possibilités. Le premier type explore l'arbre en suivant toujours l'hypothèse la plus "prometteuse", tandis que le deuxième examine parallèlement toutes les hypothèses d'un seul niveau.

2.8 Sortie d'un SRAP

Les systèmes de reconnaissance fournissent un texte représentant la transcription d'un signal sonore. Outre la transcription finale, un SRAP peut trouver les $\boldsymbol{N}$ meilleures hypothèses de transcription. Aussi, il peut être intéressé de travailler sur des sorties intermédiaires, telles que les graphes de mots et les réseaux de confusion. En plus, il est capable d'estimer la qualité de la sortie d'un SRAP en donnant l'indice de mesures de confiance. Cette section explique davantage ces différentes sorties.

2.8.1 Liste de N meilleures hypothèses

Un SRAP fournit une liste ordonnée des quelques N meilleures hypothèses trouvées pour chaque segment de parole. Cette liste présente comme intérêt d'offrir plus de richesse que les transcriptions.

2.8.2 Graphe de mots

Les graphes de mots représentent sous une forme plus ou moins compacte l'espace de recherche à un instant donné du processus multi-passes de décodage. Les graphes sont des structures dont les nœuds sont des instants du signal et où les arcs représentent des hypothèses de mots accompagnés de leur vraisemblance acoustique et de leur probabilité linguistique. En effet, les graphes de mots sont donc des objets intéressants car ils contiennent beaucoup d'informations. Ces d'informations peuvent être exploitées pour d'autres applications comme le calcul des

mesures de confiance, la recherche d'information, l'interprétation sémantique, la combinaison de systèmes etc. Cependant, ils peuvent être très gros en termes de nombre de nœuds et d'arcs et devenir difficilement manipulables. Il devient alors intéressant de les élaguer ou de les compacter, par exemple sous la forme de réseaux de confusion.

2.8.3 Réseau de confusion

Les réseaux de confusion peuvent être vus comme des graphes de mots dont certains nœuds ont été fusionnés en alignant temporellement les meilleures hypothèses issues d'un graphe de mots. La topologie d'un réseau de confusion se définit ainsi : chaque nœud correspond à un intervalle de temps et les liens entre les nœuds correspondent chacun à un mot. Ces liens sont pondérés par la probabilité a posteriori de chaque mot w, la somme des probabilités entre deux nœuds est égale à 1.

2.8.4 Mesures de confiance

Outre les hypothèses de transcription que fournit un SRAP, il est possible de calculer un scoreCM pour chaque mot. Ce score est considéré comme un indice, compris entre 0 et 1, permettant d'indiquer à quel point une décision prise par un système est fiable. Plus le score se rapprochera de 1, plus le mot est considéré correct. Ces scores sont typiquement calculés à partir des probabilités a posteriori sur les graphes de mots, les listes de N-meilleures hypothèses ou les réseaux de confusion ou d'informations dérivées des graphes de mots [Wessel et al., 2001].

2.9 Evaluation d'un SRAP

Une fois qu'une transcription est réalisée, il s'agit d'en évaluer la qualité. Une des mesures les plus répandues pour évaluer les performances d'un système de reconnaissance automatique de la parole est le taux d'erreur mot (*Word Error Rate* - WER). Le WER consiste à comparer les hypothèses de transcription et la transcription de référence. Pour ce faire, un alignement mot à mot est réalisé entre les deux transcriptions et la comparaison s'effectue selon les différents types d'erreurs sur les mots que peut commettre le système. En effet, le WER considère trois types d'erreurs :

- Insertion (I) : le système propose un mot qui n'est pas présent dans la référence ;
- Suppression (D) : le système omet un mot présent dans la référence ;
- Substitution (S) : le système remplace un mot de la référence par un autre.

Formellement, le WER est calculé comme suit :

$$WER = \frac{I + D + S}{W} * 100 \tag{2.9}$$

Où W est le nombre de mots dans la transcription de référence.

II. Aperçu sur quelques SRAP pour les langues peu dotées

À présent, la construction de corpus constitue une étape capitale pour une bonne réalisation d'outils de traitement automatique de la langue tels que les SRAP qui nécessitent une grande quantité de corpus, contenant non seulement des signaux de paroles pour l'apprentissage des modèles acoustiques du système mais également des données textuelles pour l'apprentissage des modèles de langage du système. En revanche, ces ressources sont disponibles pour certaines langues tels que le français et l'anglais nommés des langues bien dotées tandis qu'ils sont indisponibles pour d'autres langues, qui sont les langues peu dotées. Hormis, ces dernières souffrent d'un manque de données textuelles et audio en grandes quantités pour apprendre leurs vocabulaires et leurs modèles de langage. En effet, les langues peu dotées sont essentiellement orales et n'ont pas de forme écrite répandue. De ce fait, dans cette deuxième partie, nous allons définir dans un premier temps les langues peu dotées et les langues bien dotées dans le contexte du traitement de la parole, et plus précisément dans la reconnaissance de la parole. Dans un deuxième temps, nous allons passer en revue quelques exemples de SRAP pour les langues peu dotées.

À ce propos, cette partie n'est pas un catalogue pour défiler tous les SRAP pour ce type de langue, mais il met l'accent sur quelques-uns dont le point commun leur appartenance à l'ensemble des langues peu dotées. Cependant, ils se diffèrent les uns des autres au niveau de leurs origines et au niveau des solutions utilisées pour pallier le problème du manque de ressources.

2.10 Définition des langues peu dotées

En règle générale, les ressources qui sont nécessaires pour la construction d'un SRAP sont les corpus textuels, les corpus de parole et un dictionnaire de prononciation. Ces ressources sont disponibles pour des langues avec une quantité importante. Cependant pour d'autres langues ils sont indisponibles en raison du manque de ressources linguistiques.

À ce propos, [Le 2006] au cours de sa thèse intitulé « Reconnaissance automatique de la parole pour des langues peu dotées » a utilisé le critère de la disponibilité de ressources linguistiques pour classifier les langues selon deux types le premier correspond aux langues peu dotées et le deuxième type est celui des langues bien dotées.

2.10.1 Les langues bien dotées

Selon [Le 2006], une langue bien dotée est ainsi définie comme une langue possédant des ressources disponibles pour la reconnaissance automatique de la parole. Ainsi, les langues appartenant à cette catégorie sont souvent des langues très bien dotées informatiquement. Vu que l'Internet est la principale source de collecte de textes et d'audio pour constituer des corpus de taille importante, les langues bien dotées disposent d'une diffusion très large sur Internet avec une présence importante dans les médias. En effet, les langues bien dotées sont souvent des langues parlées par un nombre important de locuteurs ou encore les langues majoritaires tels que l'anglais, le français, l'espagnol, le chinois, l'arabe, l'allemand, l'italien... [Pellegrini 2008] dans sa thèse a montré que le nombre de locuteurs n'est pas un indicateur fiable de l'importance de l'utilisation d'une langue sur le Web. Ceci évoque l'exemple de la langue Quechua qui est parlée par 10 millions de personnes, et pratiquement absente sur le Web. Par contre, l'islandais est une langue parlée par 300k personnes, et elle est très dynamique et très utilisée dans les médias et sur Internet.

Par ailleurs, les langues des pays bien développés ne sont pas toujours classées parmi les langues bien dotées. En considération de ce qui précède, [Pellegrini 2008] a indiqué dans sa thèse que le niveau de développement d'un pays ne présente pas un indicateur fiable de l'importance de l'utilisation d'une langue sur le Web c'est-à-dire, une langue d'une zone développée n'est pas nécessairement une langue présentée sur Internet. Ceci fait rappel selon [Pellegrini 2008] à luxembourgeois qui dispose de très peu de visibilité sur le Web, non pas à cause du niveau de vie moyen des luxembourgeois, mais plutôt à cause de l'utilisation d'autres langues de communication dans ce pays, qui sont l'anglais, l'allemand et le français. Encore, il faut mentionner que les langues bien dotées sont souvent des langues officielles des pays. Néanmoins, le fait qu'une langue soit déclarée langue officielle n'implique pas forcément sa présence importante dans les médias et en particulier sur Internet. Tel que l'exemple de la langue Khmère qui est la langue officielle du Cambodge, parlée par une dizaine de millions de personnes dans le monde. Toutefois, elle est considérée comme une langue peu dotée du fait de la limite de ses ressources linguistiques sous forme numérique (corpus de texte et d'audio) selon [Seng 2008].

En somme, la présence importante d'une langue dans les médias et en particulier dans les télévisions, ainsi que le nombre de locuteurs et le niveau de déve-

loppement d'un pays ne présentent pas des indicateurs fiables de l'importance de l'utilisation d'une langue sur le Web.

2.10.2 Les langues peu dotées

Théoriquement, une langue peu dotée est définie comme une langue qui ne possède pas encore ou pas beaucoup (en quantité et en qualité) de ressources linguistiques pour la construction d'un système de reconnaissance automatique de la parole, particulièrement dans un contexte d'apprentissage où les données doivent être disponibles en grande quantité. Ainsi, les langues appartenant à cette catégorie sont fréquemment des langues peu dotées informatiquement.

Après avoir défini les langues peu dotées et les langues bien dotées, nous allons exposer dans la section qui suit quelques SRAP pour les langues peu dotées. De plus, nous allons exposer les stratégies suivies pour l'acquisition des ressources nécessaires pour le développement de ces SRAP et enfin nous allons monter les résultats d'évaluation de ces systèmes.

2.11 Les SRAP pour les langues peu dotées

Dans cet ordre d'idées, nous allons commencer par la première langue disposant peu de ressources électroniques et peu informatisées qui est la langue Swahili.

2.12 Un SRAP pour la langue Swahili

À une vaste région de l'afrique de l'est, le « swahili [1] » est la langue la plus répandue et la plus parlée par les communautés. Cette langue véhiculaire couvre un large territoire de presque plus de huit pays (langue nationale au Kenya et en Tanzanie) [Edgar 1967] avec un nombre compris entre 40 et 100 millions de locuteurs dont moins de 5 millions sont des locuteurs natifs selon les indications de la majorité des estimations. À priori, le Swahili se caractérise par une morphologie verbale complexe avec une absence de tons, ainsi que par une partie importante de son vocabulaire d'origine arabe. En effet, cette langue est le fruit d'un métissage de langues africaines, d'arabe et de persan. Par ailleurs, cette langue dispose d'une orthographe très proche de sa prononciation et très régulière, autrement dit, il s'agit pour chaque phonème ou une unité de base linguistique, une seule même forme écrite adoptée.

Toutefois, ce langage est considéré comme une langue peu dotée du fait de la limite de ses ressources linguistiques sous forme numérique quelque soit pour

1. http ://fr.wikipedia.org/wiki/Swahili.

les corpus du texte ou les corpus de parole [Gelas 2012]. De ce fait, des travaux antérieurs ont porté sur plusieurs méthodes de collection des données. En effet, parmi ces travaux tenons l'exemple du travail du [Gelas 2012] qui ont développé un SRAP à grand vocabulaire pour le Swahili. Dans ce qui suit, nous décrivons le processus suivi pour la construction automatique d'un corpus (audio et texte) afin de construire un modèle de langue et un modèle acoustique.

2.12.1 Recueil des ressources

Un corpus de texte est indispensable pour la modélisation du langage du SRAP. Puisque le swahili bénéficiant d'une bonne visibilité sur le web, [Gelas 2012] ont choisi de suivre l'approche qui consiste à récupérer les données existantes sur 16 sites d'informations présélectionnés pour être strictement monolingue, dans le but d'obtenir une grande quantité de textes rapidement et gratuitement. Afin d'avoir des données exploitables, ces auteurs ont été téléchargées toutes les pages d'articles d'informations sous différents formats, auxquelles ils ont appliqué les processus d'extraction de texte, nettoyage et filtrage. En conséquence, les auteurs ont récupéré plus de 28M de mots à travers ce processus.

Cependant, la richesse morphologique et l'importante de la variété lexicale de la langue Swahili a entrainé un manque de données et une mauvaise couverture lexicale pour le RAP. En général, la contrainte majeure se réside dans le taux élevé des mots hors vocabulaire (HV) qui a une grande influence sur le taux d'erreurs des mots du système. Effectivement, chaque mot HV ne sera pas reconnu mais influera aussi la reconnaissance des mots voisins avec une montée du taux d'erreur [Gelas 2012]. En revanche, ces auteurs ont recours à une solution pour pailler le problème rencontré. La solution proposée est d'élargir la couverture lexicale le plus possible en segmentant les mots en sous-unités connus sous le nom de « morphe [2] ». Conséquemment, après la segmentation en morphes effectués, ces auteurs ont obtenu à la fin une meilleure couverture lexicale tout en gardant la même taille de vocabulaire avec environ 19.17% de types hors vocabulaire avec un lexique de 65k mots et 11.36% avec 65k morphes.

En effet, il faut mentionner qu'afin d'effectuer l'apprentissage des modèles acoustiques, il est nécessaire d'avoir des données audio ainsi que les transcriptions correspondantes. Cependant, dans une situation de langues peu dotées, l'obtention et la collecte des ressources audio représente une contrainte majeure au déploiement d'un SRAP. Théoriquement, il existe de nombreuses études qui ont suggéré des méthodes afin d'accélérer la création de ces types de corpus. Certes, le travail de [Gelas 2012] consiste en premier lieu à collecter un corpus de parole lue. En

2. Le terme morphe est utilisé ici pour cette unité entre la syllabe et le mot. Selon le type de segmentation, elle peut correspondre au morphème, unité minimale porteuse de sens. Mais dans certain cas, avec une segmentation non-supervisée, elle peut ne correspondre à aucun type d'unité linguistique.

deuxième lieu, ces auteurs ont produit des enregistrements qui ont été faits par 5 locuteurs natifs (2 femmes et 3 hommes), totalisant ainsi 3 heures et demie de parole lue. Dans le but de fournir rapidement les transcriptions de ce corpus, les auteurs [Gelas 2012] ont utilisé un outil nommé crowd sourcing Amazon's MechanicalTurk (MTurk). Ce dernier est un marché de travail en ligne où quiconque peut soumettre de simples tâches à des personnes volontaires. À l'aide de cet outil, les 3 heures et demie obtenues auparavant ont été transcrites. De même, ils ont recours à collecter plus de 200h d'émissions d'informations radio diffusées sur le web afin d'obtenir un corpus plus conséquent. [Gelas 2012] ont utilisé à un processus de transcription collaboratif avec l'institut « kenyan [3] » pour transcrire 12 heures de corpus d'émissions d'information radio diffusées.

L'optique principale de la recherche effectuée par [Gelas 2012] vise à faciliter et réduire le temps de la transcription. Ainsi, un autre processus de transcription est celui de transcription collaboratif basé sur l'application itérative du protocole qui est le suivant : en premier temps un modèle acoustique est appris en utilisant les données du corpus de parole lue. En deuxième temps, un ensemble de deux heures d'audio pré-segmentées et pré-filtrées est transcrit par leur premier SRAP. Puis, la sortie de ce décodage était vérifié par des transcripteurs et finalement une version corrigée est prête. D'après la méthode proposée par [Gelas 2012], cette procédure doit être répétée jusqu'à ce que 12 heures de paroles transcrites soient obtenues, en gardant 10 heures pour l'apprentissage et 2 heures pour le corpus de test.

Généralement, une fois un corpus textuel est traité, le dictionnaire de prononciation qui représente l'élément primordial de la modélisation acoustique des SRAP doit être généré. De ce fait, la génération de ce dictionnaire était faite à travers l'extraction du corpus de texte construit disposant les 65k mots les plus fréquents. L'étape qui suit consiste à fournir une prononciation pour chacune des entrées lexicales en utilisant un nombre limité de phones qui est l'unité de base des modèles acoustiques. Etant donné que l'orthographe swahili est très proche de sa prononciation et très régulier, un script graphème vers phonème tire pleinement bénéfice de cette régularité et permet de générer la majeure partie des prononciations.

2.12.2 Expérimentations

À cet égard, après avoir collecté toutes les ressources décrites auparavant, les auteurs ont utilisé une boîte à outils Sphinx, dans le but de développer les modèles acoustiques à base de modèles de Markov cachés à 3 états pour la langue Swahili. En outre, la construction d'un SRAP a été faite tout d'abord par l'extraction des paramètres acoustiques à travers une fenêtre glissante. Aussi, il faut noter que

3. http ://www.taji-institute.com/

chaque trame possédant une taille de 25ms au début est incrémentée de 10ms. Encore, le paramétrage du signal audio est fixé selon 13 coefficients MFCC (*Mel Frequency Cepstral Coefficients*). En ce qui concerne la modélisation du langage d'une langue morphologiquement riche, l'utilisation de sous-unités aux mots a permis l'amélioration des performances de leur système, ils ont passé de 35.7% de taux d'erreurs pour le modèle de mot à 34.8 % du taux d'erreurs avec le modèle de morphes.

Traversant la corne de l'Afrique à partir du golfe d'Aden allant jusqu'aux la péninsule Arabique pour explorer leurs langages utilisés. D'une manière générale, la langue arabe, étant la langue formelle et officielle de cette péninsule Arabique, reste toujours une deuxième langue pour tous les arabophones. En réalité, l'arabe dialectal est la langue maternelle de chaque parleur arabophone et qui est la principale variété utilisée dans la vie quotidienne pour la communication oral. En effet, dans le cadre de NLP, la langue arabe se caractérise par une grande quantité de ressources pour développer un SRAP contrairement aux dialectes arabes qui représente un problème majeur pour le développement des SRAP à cause de la faible densité des ressources. Dans ce cadre des dialectes arabes nous allons présenter par la suite un exemple de SRAP pour les dialectes arabes notamment le dialecte qatarien.

2.13 Un SRAP pour le dialecte qatarien

Nous avons choisi le dialecte qatarien comme un exemple de l'arabe dialectal. Il est la langue parlé à l'État du Qatar. Assurément, ce dialecte se réfère aux langues peu dotées vu la limite de ses ressources linguistiques sous forme numérique (corpus du texte et d'audio). Ce problème du manque de ressources est justifié par le fait que cette langue est purement parlée et non pas écrite. Quoique que les ressources du dialecte qatarien soient dérisoires, il est possible de bénéficier de grandes ressources de parole de MSA et des textes existants. Dans cette optique et afin d'explorer ce dialecte nous avons recours à traiter un exemple de SRAP développé par [Elmahdy 2014]. Les auteurs ont proposé d'utiliser conjointement les données du dialecte qatarien et les données de MSA. Dans le cadre proposé, l'utilisation d'une quantité importante de données de MSA ainsi une petite quantité de données de l'arabe dialectal consiste à améliorer les modèles acoustiques et la modélisation du langage du dialecte qatarien.

Avant de traiter les résultats et les expériences, nous allons poursuivre la présentation de la vue d'ensemble du SRAP développé pour le dialecte qatarien par [Elmahdy 2014].

2.13.1 Recueil des ressources

En fait, l'élaboration du SRAP de [Elmahdy 2014] pour le dialecte qatarien a été faite par le biais de deux types de données : le corpus arabe et le corpus dialectal. Pour le corpus arabe, [Elmahdy 2014] ont tiré profit des données de paroles du domaine d'émission de nouvelles. Ce corpus a été pris de la part de l'Association européenne pour les ressources linguistiques (ELRA). Certes, les ressources de parole ELRA sont : d'une part le NEMLAR « Corpus de parole de Diffusion des Nouvelles » qui est constitué d'environ 40 heures de différentes stations de radio. D'autre part, le NetDC « Corpus de parole arabe de diffusion des Nouvelles » contient environ 22,5 heures enregistrées à partir de Radio Orient. En somme, ils ont obtenu en total 62.5 heures d'enregistrement. Pour le corpus du dialecte qatarien, ce dernier a été collecté à partir de différents programmes de séries télévisées et de talk-show. Ainsi, les données sont sélectionnées parmi les programmes dans lesquels la majorité de la parole est en dialecte qatarien. En ce qui concerne la transcription de ce corpus, les données ont été transcrites manuellement dans les normes de l'orthographe traditionnelle de l'arabe. Pour faire référence aux consonnes arabes non-standard après la transcription, il a été nécessaire d'utiliser les cinq lettres Persane à part les lettres de l'alphabet arabe.

Compte tenu de ce qui précède, le corpus du dialecte qatarien a comporté 16 heures d'enregistrement qui a été subdivisé en un ensemble de 13 heures pour l'enregistrement, un ensemble d'une heure pour le développement, et un ensemble de 2 heures pour l'évaluation.

Dans le but de bien explorer ce SRAP qatarien, nous allons décrire les expériences faites et les résultats obtenus dans la section ci-dessous.

2.13.2 Expérimentations

Le SRAP est une architecture GMM-HMM basé sur le système de reconnaissance de la parole Kaldi [Povey 2011]. Ainsi, les modèles acoustiques sont représentés en tri-phones avec trois états par HMM formé avec un maximum d'informations Estimation mutuelle (**MMIE**). Le vecteur de caractéristique est constitué des coefficients MFCC de 39 dimensions standard. Pendant l'apprentissage de modèle acoustique, l'analyse discriminante linéaire (**LDA**) et linéaire du maximum de vraisemblance transforment (**MLLT**) sont appliquées pour réduire la dimensionnalité, qui améliore la précision ainsi que la vitesse de reconnaissance. Feature-espace MLLR (**fMLLR**) a été utilisé pour l'apprentissage d'adaptation de locuteur des modèles acoustiques.

Le modèle de langage est un modèle tri-gramme back off avec modification lissage Kneser-Ney. Le modèle de langage a été formé avec les transcriptions de

l'ensemble de l'apprentissage en dialecte qatarien (65K mots). La taille du vocabulaire est d'environ 15.5K mots uniques.

L'évaluation du système de référence donné lieu WER de 61,7% sur l'ensemble du développement du dialecte qatarien et 80,8% sur l'ensemble de l'évaluation. En examinant les résultats, il a été constaté que près de 1,0% des erreurs sont causées soit par les différentes formes de « Alef » ou finale « Alef Maksura ». Comme il n'y a pas de forme orthographique standard pour l'arabe dialectal et ces types d'erreurs sont déjà variantes orthographiques communes en arabe dialectal, [Elmahdy 2014]ont décidé d'ignorer ces types d'erreurs en normalisant la fois hypothèse et de référence. Après l'application de la normalisation orthographique, WER absolue diminue à 60,9% sur l'ensemble de développement avec réduction relative de 1,3% et de 79,9% sur l'ensemble de l'évaluation avec réduction relative de 1,1%.

Pour faire face au problème du manque des données, l'idée de base de [Elmahdy 2014] est d'utiliser une approche nommée « **cross-lingual** » dans laquelle les données en MSA et les données dialectales ont été regroupées et utilisées ensemble pour former le modèle acoustique et le modèle du langage.

Dans l'objectif d'améliorer le modèle de langage, un modèle de langage de MSA de type trigramme est entraîné en utilisant le corpus «LDC Gigaword» qui se compose de plus de 800M mots. Ainsi, le vocabulaire MSA comprend 256K mots du corpus.

En appliquant le Cross-lingual MSA/dialecte qatarien modèle du langage, le WER absolue est devenu 56,0% pour l'ensemble de développement et 64,4% pour l'ensemble d'évaluation avec une réduction relative significative de 3,6% et de 16,3% par rapport au système de référence.

Initialement, le modèle acoustique de MSA est utilisé pour décoder les données de parole du dialecte qatarien. Cependant, cela n'était pas faisable à cause de l'existence de quelques différentiations au niveau des phonèmes entre le MSA et ce dialecte. Ensuite, cette incohérence a été résolue en appliquant une correspondance entre les phonèmes. Ainsi, cette idée consiste à faire la correspondance des consonnes qui n'existent pas dans MSA à tous les phonèmes possibles et les plus proches de MSA. Après avoir réalisé la correspondance des phonèmes du dialecte qatarien, un modèle acoustique graphémique MSA est formé en utilisant un corpus de MSA de taille 62,4 heures.

Tout bien considéré, les résultats de décodage WER absolue sont de 61,9% pour l'ensemble d'évaluation et de 81,3% pour l'ensemble d'évaluation avec une augmentation relative de 1,8% par rapport à la base de référence dialecte qatarien.

Dans la mutualisation des données de la modélisation acoustique, le modèle acoustique est formé d'une façon conjointe en utilisant deux données du dialecte qatarien et MSA. À ce sujet, les résultats de décodage WER absolue sont de 56,6% et de 64,4% sur l'ensemble de développement et d'évaluation définit respectivement en super formant le système de référence par une diminution relative de 7,1% et 19,4%.

Les techniques du modèle acoustique d'adaptation sont affectées sur le modèle MSA en utilisant des données de parole du dialecte qatarien. Ainsi, les résultats de décodage WER absolue sont de 57,3% et de 65,9% sur les mêmes ensembles de développement et d'évaluation définis respectivement en surclassant le système de référence par une diminution relative de 5,9% et 17,5%.

Le modèle acoustique d'adaptation est réalisé sur le modèle MSA/dialecte qatarien de mutualisation des données plutôt que sur le modèle MSA. Pour les résultats de décodage WER absolue sont de 55,6% pour l'ensemble de développement et de 62,5% pour l'ensemble d'évaluation en surperformant le système de référence par une diminution relative de 8,7% et 21,8%.

En conclusion, à cause du manque de ressources de parole dialectales en utilisant les données MSA, la correspondance de phonème dialectal, la mutualisation des données et le modèle acoustique d'adaptation ont atteint 21,3% et 28,9% en établissant une réduction par rapport WER sur l'ensemble développement du dialecte qatarien et l'ensemble d'évaluation respectivement.

2.14 Conclusion

Ce chapitre est composé de deux principales parties. Nous avons présenté dans la première partie de ce chapitre un survol sur la RAP ainsi que les enlèves des différentes théories et les composantes d'un SRAP allant de l'extraction des paramètres du signal de la parole jusqu'à sa transcription. La deuxième partie de ce chapitre s'attarde sur la thématique particulière qui nous intéresse, à savoir le développement d'un SRAP pour les langues peu dotées. Ce choix est justifié par le fait que le SRAP, que nous allons développer, est dédié au dialecte tunisien qui est considéré comme une langue peu dotée. De ce fait, nous avons consacré cette partie pour définir en premier lieu les langues peu dotées. En deuxième lieu, nous avons souligné particulièrement les stratégies suivies pour l'acquisition des ressources nécessaires pour le développement d'un SRAP pour ce type de langues. À cet égard, nous avons choisi en premier lieu un SRAP proposé pour la langue africaine, à savoir la langue Swahili. En outre, nous avons exploré en deuxième lieu un SRAP pour les dialectes arabes tels que le dialecte qatarien.

Etant donné qu'un dictionnaire de prononciation est constitué comme un élément central de l'apprentissage des modèles acoustiques dans un SRAP, plusieurs travaux dans la littérature ont surgi afin de créer ce dictionnaire. En fait, le processus de création de ce dictionnaire phonétique est nommé conversion graphème en phonème (G2P). Le chapitre 3 s'intéresse à la présentation de l'état de l'art de la conversion G2P.

Chapitre 3

Etat de l'art sur la conversion G2P

Sommaire

3.1 Introduction

Au premier abord, l'idée de base des systèmes consacrés aux écritures alphabétiques consiste à avoir la forme orthographique en tant qu'une représentation conventionnelle de la prononciation d'un mot en utilisant des signes d'écriture [**?**]. Certes, l'association entre les lettres et les sons est considérée ambiguë et dépendante du contexte dans la plupart des langages naturels. Parmi ces langages on trouve l'espagnol, le swahili et le finnois qui disposent d'une correspondance plus ou moins directe entre l'écriture alphabétique et les systèmes phonétiques utilisés. De plus, la transcription de ces langues en formes phonétiques semble une tâche traitable en se basant sur des règles phonétiques simples et dépendantes de la langue. Il existe d'autres langues qui ont seulement des régularités partielles entre leur orthographe et les systèmes phonétiques comme l'anglais et le français

ce qui engendre une ambiguïté dans la correspondance entre les systèmes orthographiques et phonétiques. À cet effet, la transcription basée uniquement sur les règles constitue une tâche difficile pour ces langues. Pour la langue arabe la correspondance entre les systèmes orthographiques et phonétiques se situe entre le simple (espagnol, finlandais et swahili) et le complexe (anglais et français) [**?**].

Étant donné que la relation entre la forme orthographique et phonétique d'un mot est d'une manière générale indirecte, plusieurs tentatives ayant l'objectif de trouver des solutions pour métamorphoser un texte donné (des graphèmes) en des symboles phonétiques prédéfinis (des phonèmes) ont surgi. Ainsi, ce type de système est nommé un système de conversion graphème en phonème (G2P). En effet, il existe un nombre relativement important de travaux de recherches proposés dans la littérature pour la conversion G2P. Ces travaux sont classés en trois catégories selon l'approche suivie et les ressources utilisées. En premier lieu, nous trouvons l'approche manuelle, qui est l'approche la plus simple, implémentée manuellement par des experts humains (des linguistes). En second lieu, on trouve l'approche à base de règles qui consiste à exploiter les connaissances linguistiques et phonétiques des spécialistes et des experts, dans le but de développer un ensemble de règles de conversion lettre à son. En dernier lieu, l'approche « guidée par les données » qui ne nécessite pas d'experts linguistiques ou de connaissances phonétiques mais seulement une grande quantité de données d'apprentissage.

Dans ce chapitre, nous allons définir, dans un premier temps, la tâche de la conversion G2P particulièrement dans le domaine de la reconnaissance automatique de la parole. Dans un deuxième lieu, nous passons en revue les approches principales et techniques existantes dans la littérature pour la conversion G2P.

3.2 La conversion G2P

La conversion graphème en phonème (G2P) constitue la tâche qui permet d'associer à chaque séquence de graphèmes une suite de phonèmes qui lui est propre. Ainsi, un phonème peut se référer à une ou plusieurs lettres différentes (graphèmes), cela signifie qu'il existe une variation de prononciation pour une séquence de graphèmes. Ainsi avoir toutes les séquences de phonèmes correspondant à une séquence de graphèmes semble une étape nécessaire.

Bien entendu, le système de conversion G2P présente une application dans nombreux domaines tels que les systèmes de la reconnaissance automatique de la parole, les systèmes de dialogue et les systèmes de synthèse tout en se basant sur sa tâche primordiale qui consiste à convertir une séquence de graphèmes en une séquence phonétique. Dans le cadre de la reconnaissance automatique, le système de conversion G2P permet de générer un dictionnaire de prononciation. Ce dernier est un élément central de l'apprentissage des modèles acoustiques. Il s'agit

d'associer chaque entrée du dictionnaire, qui est présentée sous la forme d'une séquence de graphèmes (i.e. chaque mot) à une suite de phonèmes qui lui est propre. Généralement, le développement de ce dictionnaire repose sur trois étapes fondamentales. La première étape consiste à sélectionner l'ensemble des entrées lexicales constituant le vocabulaire qui définit l'ensemble des mots qu'un SRAP est capable de reconnaître. La deuxième étape permet de définir des unités acoustiques élémentaires qui sont les phonèmes. Finalement, la troisième étape représente la description de chaque entrée lexicale en utilisant ses unités acoustiques.

3.3 Les approches de la conversion G2P

Dans la littérature, il existe trois approches qui ont été utilisées pour la conversion G2P à savoir, l'approche manuelle, l'approche à base de règles où la connaissance linguistique et phonétique des experts est utilisée pour développer un ensemble de règles et l'approche guidée par les données. De ce fait, le reste de ce chapitre est dédié à présenter ces trois approches tout en précisant les travaux qui les ont déjà traitées et en abordant les avantages et les inconvénients de chacun des types de ces approches proposées.

3.3.1 Approche manuelle

Cette approche consiste à construire d'une façon manuelle un dictionnaire phonétique d'une taille importante par des experts humains. En effet, c'est un dictionnaire de prononciation avec des entrées écrites manuellement. Nous citons à titre d'exemple le dictionnaire CMU anglais nord-américain qui a été traitée manuellement. Ce dictionnaire contient plus de 125.000 mots et leurs phonétisation. L'ensemble de phonèmes actuels contient 39 phonèmes anglais, dont les voyelles peuvent en outre porter l'accent lexical. En raison du grand nombre d'exceptions de prononciation en anglais, la construction de ce dictionnaire a pris un temps énorme, il a fallu plusieurs années aux experts pour le terminer. En effet, pour les langues morphologiquement riches, tel que MSA, la création manuelle d'un dictionnaire de prononciation est une tâche difficile vu l'importance du nombre de formes des mots, dont chacune a de nombreuses prononciations possibles.

En outre, cette approche a des atouts telles que sa technique simple et efficace aussi sa création d'un dictionnaire de bonne qualités et même sans erreur. Toutefois, elle a des inconvénients au niveau de la construction du dictionnaire de prononciation de taille importante, qui semble une tâche fastidieuse et coûteuse en termes de temps et d'experts humains.

Afin de surmonter les limites de la première approche, une approche à base de règles a été proposée. Nous présentons cette approche dans la sous-section suivante
.

3.3.2 Approche à base de règles

Le principe de cette approche est d'utiliser des règles phonétiques pour la conversion G2P. Ainsi, cette conversion selon cette approche n'exige pas une grande quantité de corpus, mais nécessite une bonne connaissance de la langue et de ses règles phonétiques.

Dans la littérature, nous pouvons trouver plusieurs travaux qui ont utilisé l'approche à base de règles pour effectuer la conversion G2P des textes dans divers langues. Nous pouvons entre autres citer le travail de [Béchet 2001] pour la langue française. En effet, ce dernier propose un système de conversion G2P complet appelé Lia-Phon [Béchet 2001] à base de règles, dans le but est de transcrire automatiquement les graphèmes en phonèmes. Pour ce faire, le système se compose de trois modules à savoir, le module de formatage et d'étiquetage dont le rôle est le traitement du texte brut pour être prêt à la conversion G2P. Le deuxième module s'intéresse à la conversion G2P. Ce module regroupe un ensemble de bases de règles de la conversion G2P relatives aux étiquettes préalablement posées et des conditions pour effectuer les liaisons. Enfin, le module d'exploitation de la conversion G2P permet d'adapter la sortie du système à l'application visée : gestion des pauses pour servir d'entrée à un synthétiseur de parole, génération de prononciations multiples pour la phonétisation de lexiques utilisés dans des SRAP, etc.

De même, [Alghamdi 2002] ont utilisé l'approche à base de règles pour la génération des phonétisations des mots diacritiques en langue arabe. Leur système emploie les règles de prononciation spécifiques de MSA ainsi que certaines particularités de l'arabe dialectal. Dans un travail similaire sur la tâche de conversion G2P, [Tebbi 2007] ont proposé un système de conversion G2P des textes diacritiques en MSA. Leur système est appelé CGPAS : Conversion des graphèmes phonèmes des textes en MSA). Ces auteurs ont montré que la majorité des erreurs de conversion G2P provenaient des noms propres et des mots d'exceptions qui ne se lisent pas suivant des règles phonétiques bien déterminées. Dans l'intention de remédier ce problème, ils ont utilisé les règles phonétiques et aussi un lexique d'exceptions. Il faut noter que ce lexique d'exception est consulté avant que les règles soient utilisées. De ce fait, si le mot figure parmi les exceptions, le système G2P génère directement une entité lexicale qui représente la prononciation qui lui correspond. Sinon, le système doit convertir chaque graphème en un, plusieurs ou zéro phonème(s) selon le contexte et cela grâce à l'utilisation des règles phonétiques. Dans le même ordre d'idées, [cal Imed jdouben 2012] ont utilisé l'approche à base de règles pour réaliser un système de conversion G2P automatique pour le MSA. La réalisation de ce système est basée sur deux méthodes différentes. La première méthode est fondée sur l'utilisation d'un lexique qui contient une liste de mots d'exceptions et des abréviations, en introduisant directement la conversion G2P correspondante aux mots sans passer par la base de règles de transcription pho-

nétique, ce qui assure plus de rapidité dans le traitement. La deuxième méthode consiste à traiter le reste du texte en utilisant une base de règles de transcription phonétique. La structure des règles élaborées est de la forme suivante : chaque graphème est remplacé par un ou plusieurs phonèmes selon son contexte gauche, son contexte droit, ou les deux à la fois.

Récemment, la plupart des travaux concernant les SRAP pour le MSA ont exploité un dictionnaire de prononciation, construit en associant chaque mot non voyellé à toutes les possibilités de voyellations de ce mot. En effet, l'apparition de techniques d'analyse morphologique et de désambiguïsation peut également être utile dans le cadre de la création automatique d'un dictionnaire phonétique. En outre, ces techniques serent à déterminer la prononciation la plus probable d'un mot selon son contexte. Etant donné que l'absence des voyelles en MSA engendre une ambigüité dans la lecture et même la sémantique d'un mot, [Vergyri 2008] propose d'utiliser un analyseur morphologique et de désambiguïsation MADA [Habash 2006]. Grâce à cet analyseur, le problème de désambiguïsation d'un mot est résolu et la prédiction de la prononciation devient simple. En effet, ils ont choisi le meilleur choix du système MADA afin de générer la prononciation d'un mot.

Jusqu'à présent, l'arabe dialectal est devenu la langue des nouvelles et de nombreuses variétés de programmes de télévision, ainsi que de la communication informelle en ligne, dans des courriels, blogs, forums de discussion, chats, SMS, etc. Pour cette raison, les dialectes arabes ont acquis le statut des langues vivantes dans les études linguistiques, nous voyons ainsi l'apparition d'un effort sérieux pour étudier des modèles et des régularités dans ces variétés linguistiques. Dans le cadre de la conversion G2P pour l'arabe dialectal, [Harrat 2014] ont utilisé l'approche à base de règles pour générer la prononciation des mots du dialecte algérien. De ce fait, afin d'éviter toute ambiguïté provoquée par l'absence des signes diacritiques, ils ont ajouté tout d'abord les signes diacritiques aux textes en utilisant un système de diacritization automatique nommée ADAD (AutomaticDiacritizer of Algerian Dialect). Puis, ils ont appliqué plusieurs règles phonétiques définies par des experts pour réaliser la transformation de G2P. La plupart de ces règles sont applicables seulement pour les mots arabes. Néanmoins, le dialecte algérien se caractérise par la présence des mots empruntés du français [Harrat 2014]. Par conséquent, la transcription phonétique de ces mots ne peut être faite en utilisant les règles phonétiques conçues pour MSA ; mais nécessite des techniques plus sophistiquées en utilisant principalement des approches statistiques.

Cette approche présente différents avantages. Tout d'abord, les systèmes à base de règles permettent de fournir des dictionnaires de prononciation de bonne qualité. Ensuite, ce type de systèmes permet de mieux contrôler la qualité de la construction des dictionnaires de prononciation c'est à dire en cas d'erreur il est possible d'ajouter une nouvelle règle. Enfin, un système fondé sur des règles ne nécessite

pas une grande quantité de corpus, mais nécessite une bonne connaissance de la langue et de ses règles phonétiques.

Cependant, les problèmes auxquels sont confrontés les systèmes de conversion G2P à base de règles sont nombreux. Dans cette section, nous avons cité les problèmes les plus connus. Premièrement, l'élaboration des règles phonétique est difficile et nécessite des compétences linguistiques et spécifiques de cette langue. En effet, cette tâche exige un grand nombre d'expertise. Deuxièmement, les langues naturelles présentent souvent des irrégularités, qui doivent être capturés par les règles d'exception ou des listes d'exceptions. Dans le même ordre d'idées, l'interdépendance entre les règles peut être très complexe, les concepteurs de règles doivent vérifier avec des contres exemples si le résultat de l'application des règles est correct dans tous les cas. Cela rend le développement et la maintenance de systèmes de règles très pénibles dans la pratique. Enfin, le système G2P à base de règles est toujours susceptible de commettre des erreurs en cas de présentation d'un mot exceptionnel qui n'est pas pris en considération par le concepteur des règles.

3.3.3 Approche guidée par les données

Contrairement à l'approche à base de règles décrites auparavant, l'approche guidée par les données est basée sur l'idée d'avoir suffisamment d'exemples donnés pour la prédiction la prononciation des mots non vues purement par analogie. En fait, cette approche possède des points forts dans le fait qu'elle ne nécessite pas des experts hautement qualifiés pour la construction manuelle des règles. De surcroît, pour les locuteurs natifs, il est plus facile de juger l'exactitude de la prononciation ou d'écrire la prononciation d'un mot spécifique que de formuler des règles d'orthographe générales.

En revanche, l'approche guidée par les données ne fonctionne bien que dans le cas de présence d'une grande quantité de corpus d'apprentissage pour atteindre les performances du niveau d'un système à base de règles bien conçues ; en utilisant uniquement les données. Plusieurs méthodes de l'approche guidée par les données ont été suggérées dans la littérature, y compris la prononciation par analogie, classification locale (réseaux neuronaux, arbres de décision), HMM, JMM, CRF et Hidden CRF.

3.3.3.1 Les techniques basées sur la classification locale

Dans une approche guidée par les données, la plupart des méthodes de G2P nécessitent tout d'abord un alignement entre les graphèmes et les phonèmes du corpus d'apprentissage. L'alignement est fait dans une phase de prétraitement séparé. Durant cette phase, chaque élément d'alignement comprend un seule graphème,

zéro ou une ou même plusieurs phonèmes qui lui correspondent. Pour cette raison, le symbole « - » (absence de lettre, absence de phonème) est introduit dans l'alignement, du côté des lettres et des phonèmes. Généralement, ces alignements peuvent être réalisés soit à travers des règles conçues manuellement, par programmation dynamique basée sur des contraintes prédéfinies ou même par une estimation itérative des probabilités d'alignement. Typiquement, les graphèmes sont traités d'une manière séquentielle (par exemple de gauche à droite). La prédiction du phonème (ou parfois d'un groupe de phonèmes) est basée sur le contexte actuel du graphème. Étant donné que la décision pour chaque position est prise avant de procéder à la prochaine, nous appelons cette famille de techniques de locale classification. Après cette phase d'alignement, une technique d'apprentissage est utilisée pour permettre de prendre des décisions sur les mots non phonétisés [Laurent 2010]. Les techniques les plus couramment utilisées pour faire cette prédiction sont les réseaux de neurones et les arbres de décision. La prise de décisions de chaque phonème localement n'est pas clairement optimale à partir d'une décision au point de vue théorique. Cependant, cette stratégie permet d'éviter la nécessité d'utiliser un algorithme de recherche qui est généralement nécessaire pour trouver une solution globalement optimale.

- **Les réseaux de neurones**

Les réseaux de neurones ont un très grand pouvoir de discrimination permet de distinguer des phonèmes ayant des comportements acoustiques très proches. Leur principal désavantage est qu'ils nécessitent un apprentissage « supervisé » avec un volume de données très important. Par la suite, nous allons présenter quelques travaux qui ont utilisé ce type de technique pour accomplir la tâche de conversion G2P.

Bien entendu, [Sejnowski 1987] ainsi que [**?**] ont appliqué les réseaux de neurones à ce problème de classification. Ils utilisent un neuronal à trois couches réseau. La couche d'entrée est une fenêtre de contexte de plus ou moins trois lettres. La couche d'entrée utilise une représentation orthogonale c'est-à-dire une entrée pour chaque type de lettre. La couche de sortie représente le phonème prédit grâce à des caractéristiques articulatoires. [Jensen 2000] et [Häkkinen 2003] ont amélioré l'approche précédente en employant une représentation plus sophistiquées lettre de code-book dans la couche d'entrée. De même, [Seng 2011] ont utilisé les réseaux de neurones pour convertir les lettres en phonèmes. Le modèle de conversion est subdivisé en deux principales étapes. Le réseau de neurones de la première étape permet d'aligner la séquence de lettres M en entrée en N phonèmes en sortie en combinant deux modèles contextuels, un pour les lettres nommé LCD (LetterContext-Dependent) et l'autre concerne les phonèmes connu par PCD (PhonemeContext-Dependent). Ensuite le réseau de neurones de la deuxième étape

peut prédire le phonème final de sortie par l'observation d'une combinaison de plusieurs suites de phonèmes consécutifs celui obtenu par la première étape.

Dans le but d'améliorer les performances de la conversion G2P, le travail de [Seng 2014] traite le problème des phonèmes contradictoires, où un graphème d'entrée peut produire beaucoup de phonèmes de sortie possibles en même temps et dans le même contexte. À cette fin, ils ont proposé une approche en deux étapes à base de réseau neuronal qui convertit le texte en entrée à des séquences de phonèmes dans la première étape. Par contre dans la deuxième étape, elle prédit chaque phonème de sortie en utilisant les informations phonémiques obtenues. Ainsi, le réseau de neurones de la première étape est essentiellement mis en œuvre en tant que modèle de correspondance de plusieurs-à-plusieurs pour la conversion automatique d'un mot en une séquence de phonèmes. Tandis que la deuxième étape utilise une combinaison des séquences de phonèmes obtenues pour prédire le phonème de sortie correspondant à chaque graphème d'entrée d'un mot donné.

- **Les arbres de décision**

Les arbres de décisions donnent de bonnes performances pour les mots vérifiant les règles standards de prononciation de la langue, les résultats se dégradent rapidement en présence de mots atypiques (dont la graphie ne permet pas, en utilisant les conventions de la langue, de déterminer la prononciation).

Pour la conversion G2P, les graphèmes et leurs contextes représentent l'entrée du classificateur d'arbre de décision. Tandis que, les phonèmes représentent la sortie. Le processus de phonétisation se déroule comme suit : ***i)*** passer séquentiellement par chaque graphème et son contexte dans l'arborescence, ***ii)*** affecter à chaque classe « graphème et son contexte » un phonème, ***iii)*** et enfin concaténer ces phonèmes afin d'avoir la phonétisation finale. Afin de générer la forme phonétique de chaque séquence graphèmique, l'arbre de décision nécessite un alignement de type un-à-un entre les graphèmes et les phonèmes. Afin de répondre à cette exigence, une valeur «NULL» est inséré dans la chaîne de phonèmes dans le cas où un lettre ne correspondant pas à un phonème et vise versa.

Plusieurs travaux ont utilisé cette technique que nous allons citer quelque'uns. Le concept d'arbres de décision a été utilisé dans plusieurs travaux concernant la G2P tels que [Torkkola 1993] qui utilise une technique appelée dynamique expansion contexte qui permet de générer un arbre de décision avec l'utilisation d'une fenêtre asymétrique autour du courant lettre en considération. [Daelemans 1996] dans son travail proposent l'utilisation d'arbres de décision en utilisant le critère de gain d'information (IG-Tree). Les questions ne sont employées qu'à propos des lettres entourant et le gain d'information est calculé qu'une seule fois pour

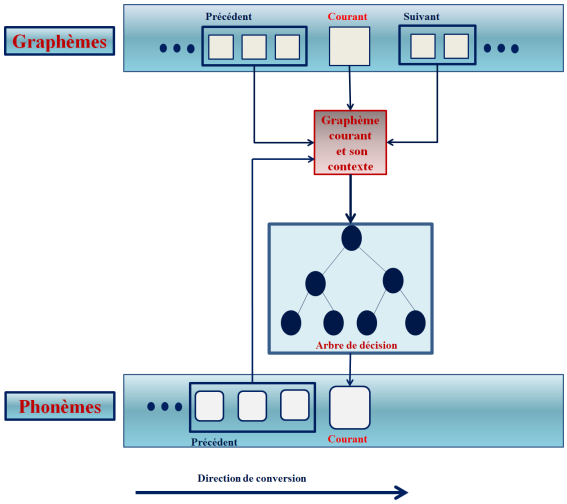

Figure 3.1 – Représentation schématique de la conversion G2P en utilisant l'arbre de décision.

chaque attribut. De même, [Andersen 1996] ont exploité le critère de Gini pour la construction d'un arbre de décision binaire. La particularité de son travail réside dans l'utilisation en même temps de cinq lettres qui se positionnent à gauche et à droite de la lettre courante. Ils ont admis aussi la possibilité de poser des questions sur l'adhésion d'un graphème voisin à une des 10 classes présentant le type d'une lettre. [Pagel 1998] ont tenté à progresser les arbres de décision à l'aide du critère de gain d'information, en calculant le gain d'information pour chaque groupe de nœud. En plus les trois lettres précédentes et suivantes ont permis également à l'algorithme de prendre les trois phonèmes suivants en compte. Ceci exige que le mot doit être traité dans l'ordre inverse de droite à gauche, depuis les phonèmes sont considérés comme le résultat de décisions prises antérieurement. Ils ont aussi signalé des améliorations par l'ajout des questions sur la partie du discours (POS) du mot considéré. Dans le même ordre des idées, [Suontausta 2000] et [Häkkinen 2003] utilisent également le gain d'information provenant des arbres de décision. La série de questions possibles comprend jusqu'à à quatre précédentes et quatre lettres suivantes ainsi que les phonèmes précédentes et leurs classes de phonèmes. Dans le travail de [Beaufort et al., 2006] la conversion graphème en phonème se déroule en deux étapes. Dans la première étape, la transcription est générée indépendamment du contexte, à l'aide d'un arbre de décision compilé à partir d'un dictionnaire d'apprentissage. L'arbre détermine pour chaque graphème le phonème qui lui correspond, en tenant compte du contexte et de la nature lexicale du mot. Si la transcription générée est incorrecte pour la langue, le mot est alors épelé. Dans la deuxième étape, la transcription du mot est remise en contexte. La prononciation d'un mot est en effet influencée par les mots qui l'entourent. Il

est dès lors nécessaire d'examiner ces mots de manière à déterminer s'ils influent sur la prononciation du mot courant.

Plus récemment, le concept d'arbre de décision binaire a été aussi exploité par [Loots 2011] pour la construction d'un système de conversion G2P. Le système a reposé sur l'idée que chaque nœud de l'arbre possède une question dont la réponse est de type vrai/faux pour chaque graphème en entrée ainsi que ses contextes sachant que la question ayant le plus grand gain d'entropie est affectée à ce nœud. Ensuite, l'arbre est parcouru de la racine en utilisant la réponse de la question de chaque nœud pour déterminer le nœud enfant à choisir. Le résultat réside dans le phonème de sortie associé à la feuille du nœud. Ce processus produit une seule prononciation de l'orthographe présentée.

3.3.3.2 Prononciation par analogie : PPA

L'idée principale de la prononciation par analogie (PPA) consiste à scanner les mots ou parties de mots existant dans le lexique d'apprentissage afin de les réutiliser pour phonétiser les mots jamais rencontrés, en utilisant une mesure de similarité entre les mots. En effet, la prononciation de l'input est alors choisie pour être analogue à des exemples récupérés. D'une façon générale, le processus du PPA se déroule sur trois phases. Dans la première phase, une correspondance entre les sous chaînes de l'input et celles des mots du dictionnaire est faite. Au cours de la deuxième phase, un treillis de prononciation est construit. En effet, il s'agit d'ajouter l'orthographe phonétique de la chaîne correspondant, à un réseau contenant toutes les prononciations possibles ainsi que les lettres de l'input qu'elle correspond. À la fin, il consiste à décider lequel des prononciations possibles est susceptible d'être le plus précis.

Cette méthode de PPA est connue par deux techniques différentes à savoir, «partial pattern matching» et «full pattern matching». La première technique consiste à comparer le premier caractère de mot d'entrée par le premier caractère de l'entrée dans le dictionnaire. À l'opposé de ce qui précède, la deuxième technique consiste à commencer la comparaison par le début de la chaîne d'entrée du dictionnaire aligné avec l'extrémité de la chaîne de mot d'entrée. Le processus est continué jusqu'à ce que l'extrémité de l'entrée de dictionnaire est alignée avec le début du mot d'entrée.

Plusieurs travaux ont utilisé cette technique que nous allons citer quelque'uns. Par exemple, [Dedina 1991] utilisent une méthode nommé «partial pattern matching». À ce propos, cette méthode consiste à examiner chaque mot dans le lexique et construire une structure en treillis de prononciation en utilisant les représentations phonétiques des mots qui correspondent à la chaîne d'entrée. Dans ce réseau de prononciation, chaque nœud représente un candidat de phonème, et chaque

chemin à travers le réseau représente une prononciation possible. Un peu plus tard et dans le cadre de la conversion G2P, [Damper 1997] ont utilisé ce qui est connu « full pattern matching». Cette méthode consiste à commencer la comparaison par le début de la chaîne d'entrée du dictionnaire aligné avec l'extrémité de la chaîne de mot d'entrée. Le processus est continué jusqu'à ce que l'extrémité de l'entrée de dictionnaire est alignée avec le début du mot d'entrée. Dans le même ordre des idées, [Marchand 2000] ont utilisé la méthode «full pattern matching». Cependant, à l'opposé de celle définit par [Damper 1997], l'idée est de commencer du début du mot d'entrée aligné avec l'extrémité de l'entrée du dictionnaire. Le processus se termine lorsque l'extrémité du mot d'entrée est alignée avec le début de l'entrée du dictionnaire.

3.3.3.3 Les approches probabilistes

Les approches probabilistes ont marqué une nouvelle aire dans l'histoire de la conversion G2P. En effet, ces approches exigent un volume important de données pour l'apprentissage. Son but est de prendre en compte des événements imprévus qui auront une répercussion moindre sur la reconnaissance. Parmi ces approches, nous pouvons citer, les Modèle de Markov caché (HMM), les modèles de séquence conjointe et les CRF. Les trois sections suivantes sont justement dédiées à la description de ces approches utilisées pour la conversion G2P.

- **Modèle de Markov caché**

Les Modèles de Markov cachés (HMM) sont des modèles statistiques utiles pour les données séquentielles. Un HMM est défini par une structure composée d'états, de transitions et par un ensemble de distribution de probabilité sur les transitions.

[Taylor 2005] a proposé l'utilisation du HMM pour la conversion G2P, dont les graphèmes sont considérés comme les observations, alors que les phonèmes représentent les états cachés. Ainsi, les transitions entre les phonèmes décrivent la probabilité qu'un phonème suivra autre. Selon [Taylor 2005] l'avantage de l'utilisation du HMM dans le G2P est d'avoir dans une même étape : alignement des graphèmes avec les phonèmes (type d'alignement 1-à-1 ou 1-à-m), estimation des probabilités de transition entre les états et estimation des probabilités d'observation qui génèrent les graphèmes de chaque état de phonème. Dans le même ordre d'idées, [**?**] ont réalisé un système de conversion G2P basée sur les HMM. La particularité de son travail réside dans l'utilisation des données acoustiques afin de définir la relation probabiliste entre graphèmes et phonèmes. En effet, ils ont utilisés un système de Kullback-Leibler (KL-HMM) basé sur les modèles de HMM. Les informations de cette relation probabiliste sont intégrées avec les informations de

séquence dans la transcription orthographique pour déduire la séquence de phonèmes ou le modèle de prononciation.

- **Modèle de séquence conjointe**

Une contribution principale dans le domaine de la conversion G2P est donnée par [Bisani 2008], où les modèles de séquence conjointe sont appliqués. L'idée de base est fondée sur le concept d'une graphone $\boldsymbol{q}$, qui est une paire d'une séquence de graphèmes f_q et une séquence de phonèmes e_q, $q = (f_q, e_s)$. Ainsi, une séquence de graphones est générée pour chaque mot suivant sa forme orthographique et sa prononciation. En général il y a plusieurs séquences de graphones qui représentent tous la même paire de graphèmes et de phonèmes, c'est à-dire pour chaque alignement possible d'une séquence de graphèmes et sa séquence de phonèmes correspondant y est une séquence de graphone unique. Étant donné un graphone est une paire de séquences, plusieurs graphèmes peuvent être alignés à plusieurs phonèmes, ce qui implique un alignement plusieurs-à-plusieurs (m-à-n). Par exemple, le mot «signe» et sa prononciation [sie] peuvent être alignés de plusieurs façons présentées dans les deux figures suivantes.

		Graphones			
Mot	**Signe**	**s**	**i**	**gn**	**e**
Prononciation	**siɲe**	**s**	**i**	**ɲ**	**e**

Figure 3.2 – Exemple 1 Alignement d'une paire de séquence graphème-phonème en utilisant le modèle de séquence conjointe.

		Graphones				
Mot	**Signe**	**s**	**i**	**g**	**n**	**e**
Prononciation	**siɲe**	**s**	**i**	**-**	**ɲ**	**e**

Figure 3.3 – Exemple 2 Alignement d'une paire de séquence graphème-phonème en utilisant le modèle de séquence conjointe.

La probabilité conjointe $p(f, e)$ de la séquence de graphèmes $\boldsymbol{f}$ et de la séquence de phonèmes $\boldsymbol{e}$ est déterminée en additionnant tous les alignements correspondants, c'est-à-dire sur tous les séquences de graphone $q_1^k \in A(f, e)$, où $A(f, e)$ est l'ensemble de toutes les séquences de graphone qui donnent $\boldsymbol{f}$ respectivement $\boldsymbol{e}$ lorsque leurs éléments sont concaténés. Ainsi, la probabilité conjointe $p(f, e)$ a

été réduite à la distribution de probabilité $p(q_1^k)$ sur des séquences de graphones correspondant q_1^k qui est modélisée en utilisant une approximation de m-grammes :

$$p(f,e) = \sum_{q_1^k \in A(f,e)} p(q_1^k) = \sum_{q_1^k \in A(f,e)} \Pi_{k=1}^{k+1}(q_k|q_{k-1},...,q_{k-M+1}) \tag{3.1}$$

Les paramètres du modèle sont estimés par la maximisation de la log-vraisemblance des données d'entraînement en utilisant l'algorithme d'espérance de maximisation (**EM**), leur initialisation est une distribution uniforme sur toute graphones $\boldsymbol{q}$ avec $|f_q| \leq L$ et $|e_q| \leq L$ pour une contrainte de longueur $\boldsymbol{L}$. La taille de la m-gramme et la limite supérieure $\boldsymbol{L}$ sont les paramètres externes du modèle de séquence conjointe et déterminent la taille de contexte, c'est à dire le nombre de graphèmes et les phonèmes qui affectent les probabilités estimées à une position donnée.

- **Conditional Random Fields : CRF**

Au cours des dernières années, CRF a reçu beaucoup d'intérêt dans de nombreuses tâches de traitement des langues naturelles qui peuvent être formulées comme des tâches d'étiquetage de séquence, comme l'étiquetage morpho-syntaxique (POS), segmentation (chunking) et la modélisation du langage. Il s'agit d'un modèle conditionnel émergé dans le cadre de modélisation discriminative. En effet, ces domaines CRF peuvent être facilement appliqués dans toutes les tâches de traduction monotone chaîne-à-chaîne, où la disponibilité d'un alignement 1 à 1 entre la source et l'emplacement ciblé est mise au point. Ainsi, dans le contexte de CRF, la conversion graphème-phonème (G2P) représente une telle tâche. C'est-à-dire, une séquence de phonèmes illustrant une prononciation valide est générée pour une séquence de graphèmes donnée. Nous présentons ci-dessous les bases mathématiques du modèle CRF dans le cadre de la conversion G2P.

Théoriquement, la tâche de la conversion G2P selon CRF consiste à trouver la plus susceptible prononciation pour un mot donné en entrée. Etant donné la séquence graphème, $f_1^J = f_1...f_j...f_J$,dont chaque f_j représente un symbole d'un alphabet d'entrée V_f.Il faut trouver la séquence de phonèmes $e_1^I = e_1...e_i...e_I$, où chaque e_i est un symbole d'un alphabet de sortie V_e, avec la plus forte véritable probabilité $Pr(e_1^I|f_1^I)$.Ainsi, la règle de la chaîne peut être appliquée pour décomposer $Pr(e_1^I|f_1^I)$ de la manière suivante :

$$Pr(e_1^I|f_1^I) = \prod_{i=1}^{I} Pr(e_i|e_1^{i-1}, f_1^J) \tag{3.2}$$

D'autant plus que généralement la probabilité $Pr(e_i|e_1^{i-1}, f_1^J)$ est inconnue, elle doit être approximée en se basant sur le calcul d'un modèle de probabilité $p(e_i|e_1^{i-1}, f_1^J)$:

$$Pr(e_i|e_1^{i-1}, f_1^J) = p(e_i|e_1^{i-1}, f_1^J) \tag{3.3}$$

En dépit de la dépendance de la probabilité d'un phonème e_i de l'ensemble de la séquence de graphèmes f_1^J et des phonèmes précédents e_1^{i-1} dans un tel modèle, celle-ci n'est pas conditionnée à des graphèmes particuliers. Il faut indiquer que cela s'explique par le fait qu'il n'existe pas encore un moyen pour exprimer toutes les relations entre les graphèmes et phonèmes représentant leurs prononciations. Par conséquent, quand formés un certain ensemble de mots et leurs prononciations, le modèle ne sera pas en mesure d'apprendre la probabilité d'un phonème conditionné sur une séquence de graphèmes, mais seulement sa probabilité conditionnée par le mot entier. En d'autres termes, ce modèle ne sera pas en mesure d'apprendre la prononciation de sous-séquences de graphèmes individuels, mais seulement la prononciation des mots entiers. Par conséquent, lorsqu'il est appliqué à prédire la prononciation des mots qui n'ont pas eu lieu pendant le processus de formation pour le modèle, il s'exécute d'une manière incorrecte.

En vue de bien clarifier les idées, nous allons présenter dans ce qui suit un exemple illustratif. Dans cet exemple nous supposons que deux paires de graphèmes-phonèmes qui sont respectivement «présent» et « pesen » apparaissent dans l'ensemble des données d'apprentissage. Du fait que la probabilité de chaque phonème est conditionnée sur la totalité de la séquence de graphème, par exemple dans les deux mots «représentation» et «représente», le modèle n'est pas capable d'apprendre la prononciation de leur tige de mot commun qui est «présent», bien que la séquence de graphèmes « présent » se produise plusieurs fois dans les données d'apprentissage.

Mot 1	Représente
Prononciation 1	ʁepʁesente
Mot 2	Représentation
Prononciation 2	ʁepʁesentasjɔ̃

Figure 3.4 – Exemple de paires de mots non alignés en terme de graphème-phonème.

Mot 3	Présent
Prononciation 3	pʁesent

Figure 3.5 – Exemple de tige de mot non aligné en terme de graphème-phonème.

Afin d'améliorer le modèle, les dépendances entre les graphèmes et les phonèmes vont être intégrées. Autrement dit, il va y avoir un moyen pour relier les phonèmes et les graphèmes qui sont liés par la prononciation. Par voie de conséquence, un alignement $a_1^J = a_1, ..., a_j, ..., a_J$ est introduit en tant que variable cachée, où $a_j = i \in \{1, ..., I\}$, si le graphème f_j génère le phonème e_i . En définitive, nous obtenons :

$$Pr(e_1^I|f_1^J) = \sum_{a_1^J} Pr(e_1^I, a_1^J|f_1^J) \tag{3.4}$$

Ainsi, en vue d'établir une relation entre les graphèmes et les particuliers phonèmes générés par eux, il faut introduire un alignement dans notre modèle. En donnant l'autorisation au modèle pour apprendre les probabilités conditionnées sur les relations entre les graphèmes et les phonèmes liés par la prononciation, permet d'améliorer la précision de la déduction de la prononciation des mots qui sont invisibles dans la formation. La raison pour cela est que le modèle n'a pas besoin pour à apprendre la prononciation de mots entiers. Plutôt, il est capable d'apprendre les probabilités des séquences de graphèmes qui génèrent des phonèmes particuliers pour un certain contexte de graphèmes et phonèmes entourant. Par conséquent, il doit connaître explicitement les phonèmes qui sont générés par quels graphèmes.

En fait, vu que l'annotation manuelle de milliers de mots semble une tâche coûteuse et fastidieuse, la plupart des corpus ne disposent pas d'alignements, d'où la question qui se pose est comment mettre en place un tel alignement pour améliorer la précision de la déduction de la prononciation des mots qui sont invisibles dans la formation. Dans ce qui suit, nous passons en revue les travaux principaux existant dans la littérature qui ont eu recours à utiliser le modèle CRF tout en exposant leurs résolutions pour les problèmes d'alignement graphème-phonème.

Selon [Dong 2011], la tâche de conversion G2P peut être considérée comme une procédure d'étiquetage par lequel l'orthographe d'un mot (de séquence graphème) est observée et la prononciation (séquence de phonèmes) est l'étiquette qui doit être déduite. En fait, en vue de former un modèle de CRF, des exemplaires étiquetés sont nécessaires. Néanmoins, la séquence de phonèmes et graphèmes d'un

mot sont habituellement de longueurs différentes, ce qui provoque un problème d'alignement. Par conséquent, [Dong 2011] a employé un JMM « joint multigram model » afin d'effectuer cette tâche d'alignement. Ce modèle JMM illustre la distribution de probabilité sur des séquences d'unités mixtes de phonèmes-graphèmes appelée graphone comme indiqué par [Bisani 2008] où 0-1 graphone utilisée pour l'alignement, ce qui signifie qu'un ou zéro phonème est autorisé à être aligné sur un graphème, et vice versa.

Dans le même ordre d'idée, [Wang 2013] ont proposé une nouvelle solution pour remédier au problème de la conversion G2P. En effet, [Wang 2013] ont tenté de présenter un système hybride combinant le modèle JMM et le CRF pour résoudre le problème de conversion G2P. De surcroît, ces chercheurs ont choisi la combinaison de ces deux modèles du fait qu'en premier lieu, le modèle JMM est capable de modéliser plus d'informations contextuelles phonétiques en raison de son modèle de langage génératif pour les n-grammes des unités lettres conjointes-phonèmes. Aussi, l'idée fondamentale de ce modèle consiste à modéliser la probabilité conjointe de la lettre et les séquences de phonèmes à la fois en considérant tous les alignements possibles lettres-phonèmes. Cependant, il faut mentionner qu'il est difficile à JMM d'intégrer des caractéristiques complexes, tels que les structures de syllabation. En deuxième lieu, les classificateurs CRF peuvent servir à effectuer la conversion G2P par la formulation de la tâche comme étant un problème de séquence d'étiquetage.

Un autre travail similaire est celui de [Illina 2011] qui ont proposé l'utilisation d'une méthode probabiliste : Conditionnelle Random Fields (CRF) pour effectuer la conversion G2P. Afin de se servir de CRF quelques prétraitements ont été réalisés. Il s'agit d'aligner les graphèmes et les phonèmes de tous les mots du corpus. Un alignement 1-à-1 est nécessaire. Afin de répondre à cette exigence, [Illina 2011] ont employé les HMM en premier lieu. Ces HMM permet de prendre facilement en compte les associations 1-à-plusieurs (un phonème à un ou plusieurs lettres). Dans son travail, chaque phonème est modélisé par un HMM discrète d'un seul Etat, chaque observation de cette HMM correspond à graphèmes. L'apprentissage de ces HMM est effectuée en utilisant intégré algorithme de Baum-Welch et la partie d'apprentissage du corpus. Après la formation, un alignement forcé de type 1-à-plusieurs entre graphèmes et phonèmes de tous les mots du corpus d'apprentissage est effectué. En second lieu, puisque les CRF exigent un alignement de type 1-à-1, ils ont extrait les associations entre une lettre et un phonème d'après l'alignement obtenu précédemment. Dans le cas où un phonème est aligné avec plusieurs lettres, ce phonème est associé à la lettre qui a la plus grande probabilité. Les lettres restantes sont associées à des phonèmes nuls "-". Après l'alignement 1-à-1 obtenu, la conversion G2P à base du modèle CRF sera formée en utilisant toutes les données de formation alignées. Un exemple illustratif est donné dans la figure 3.6 pour présenter la procédure de conversion G2P selon [Illina 2011].

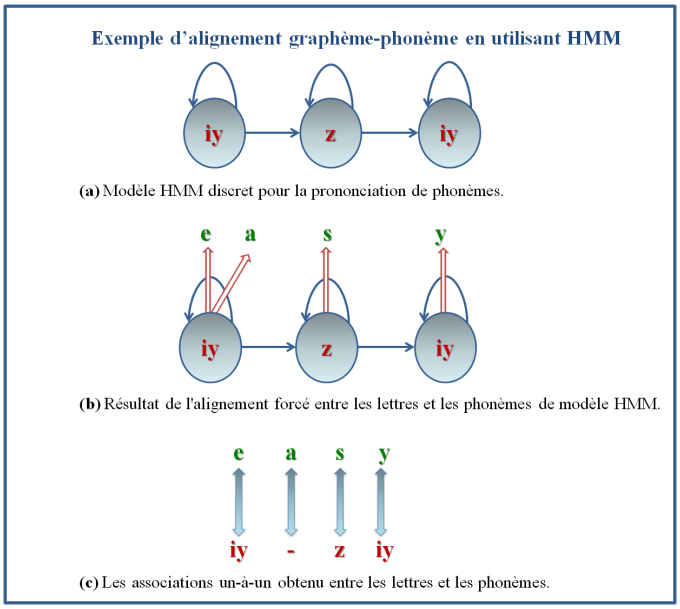

FIGURE 3.6 – Exemple d'alignement Graphème-Phonème en utilisant HMM.

Dans la même direction, [Lehnen 2011] ont utilisé CRF pour la tâche de conversion G2P. Parallèlement, ils ont trouvé une autre solution pour effectuer l'alignement de type 1-à-1 entre les graphèmes et les phonèmes en utilisant l'outil GIZA++. Ce dernier permet de traiter l'ensemble des mots en tant qu'une langue source et l'ensemble des prononciations comme une langue cible [Franz 2003]. Ce qui implique la modélisation de l'apprentissage de la correspondance entre ces deux langues sous la forme d'un problème de traduction statistique. Une fois l'alignement est effectué, ils sont passés à l'étape de formation et de test à l'aide de l'outil CRF++.

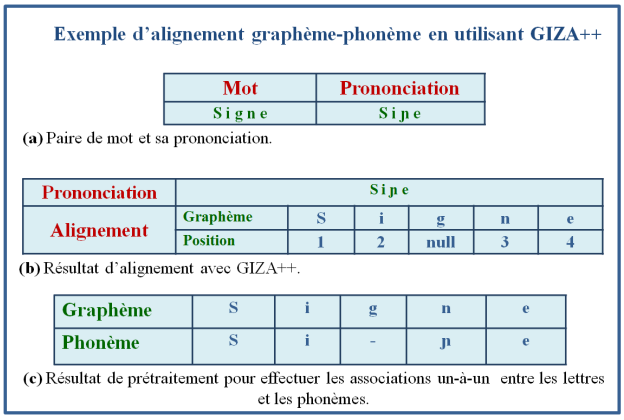

FIGURE 3.7 – Exemple d'alignement Graphème-Phonème en utilisant GIZA++.

Les travaux antérieurs ont démontré que le modèle CRF a approuvé son efficacité dans la tâche de conversion G2P, néanmoins il présente quelques limites. En effet, l'inconvénient majeur de CRF réside dans le fait que dans l'étape de formation, un alignement de type 1-à-1 est nécessaire entre les graphèmes et les phonèmes. La qualité du résultat de conversion G2P dépend fortement de cet alignement. Étant donné que ces alignements ne sont pas présentés dans le corpus annoté, des modèles externes doivent être utilisés pour produire un tel alignement dans une étape de prétraitement. Afin de résoudre ce problème, HCRF (HiddenConditional-Random Fields) sont proposés. Dans la sous-section suivante nous allons présenter davantage ce modèle.

- **Hidden Conditional Random Fields : HCRF**

Comme nous l'avons mentionné précédement pour le modèle CRF, la question qui est souvent négligée est celle comment faire face aux tâches lorsqu'il n'existe aucun alignement fourni avec les données d'apprentissage entre la source et le côté de la cible. Ainsi la réponse à cette question se trouve dans le nouveau modèle HCRF qui traite les correspondances arbitraires entre les chaînes de graphèmes et phonèmes observées dans les données d'apprentissage. Ces unités possédant des longueurs variables sont intégrées directement dans le modèle, cela permet de saisir des dépendances plus compliquées. En outre, en prenant en considération une table de conversion pré-calculée, le coût de calcul est considérablement réduit. [Lehnen 2011] ont utilisés ce type de modèle pour effectuer la tâche de conversion G2P. Nous présentons ci-dessous les bases mathématiques du modèle HCRF :

Un modèle discriminant pour la classification basée sur CRF avec variables latentes, nommé Hidden Conditionnelle Random Fields (***HCRF***), est présentée dans [Quattoni 2007]. La classification peut être décrite comme la tâche de prédiction d'une séquence de phonèmes e pour une séquence de graphèmes f_1^j . En outre, a_1^jsont les variables latentes non observées dans les exemples de formation, où chaque a_j $\boldsymbol{A}$ et $\boldsymbol{A}$ est un ensemble de phonèmes cachés possibles dans le modèle fini. Le modèle log-linéaire est de la forme :

$$p_\Lambda(e|f_1^j) = \sum_{a_1^j} p_\Lambda(e, a_1^j|f_1^j) = \frac{\sum_{a_1^j} \exp(H_\Lambda(e, a_1^j, f_1^j))}{\sum_{\widetilde{e}} \sum_{\widetilde{a_1^j}} \exp(H_\Lambda(\widetilde{e}, \widetilde{a}_1^j, f_1^j))} \tag{3.5}$$

Où Λ sont les paramètres du modèle et $H_\Lambda(e, a_1^j, f_1^j)$ est une fonction de potentiel paramétré par Λ. Pour l'estimation des paramètres du modèle, la log-vraisemblance

des données de formation, à laquelle une gaussienne avant est ajoutée pour éviter sur-apprentissage, est maximisée en utilisant l'algorithme de gradient remontée. En raison des variables latentes, le problème d'optimisation n'est pas convexe.

3.4 Conclusion

Dans ce chapitre nous avons présenté les différentes approches existantes pour la conversion G2P. Au premier abord, nous avons commencé par présenter la définition de la tâche de conversion G2P, tout en mettant l'accent sur son utilité ainsi que ses domaines d'applications. Ensuite, nous avons classé les méthodes de conversion G2P en trois grandes approches à savoir l'approche manuelle, l'approche à base de règles et l'approche guidée par les données, tout en présentant les travaux qui ont utilisé chacun des types des trois approches proposées et en abordant leurs avantages et inconvénients.

Bien que les travaux que nous avons exposés soient faits pour plusieurs langues, nous nous sommes concentrés plus sur la langue arabe vu que notre travail porte sur le dialecte tunisien grâce au grande ressemblance entre l'arabe et le dialecte tunisien. Tout bien considéré, afin de réaliser la tâche de conversion G2P de n'importe quelles langues, il est nécessaire de collecter des corpus et d'avoir une idée sur ses caractéristiques linguistiques et plus précisément ses caractéristiques phonétiques notamment si nous allons choisir l'approche à base de règles. En particulier, il semble intéressant de répondre à ses nécessités mentionnés avant de commencer le processus de conversion G2P, ce qui est l'objectif du chapitre suivant.

Chapitre 4

Recueil des corpus pour le dialecte tunisien

Sommaire

4.1 Introduction

Bien que le MSA soit la langue d'usage officiel des médias et de l'éducation, l'arabe dialectal est la langue de la vie quotidienne et la vraie forme native de l'arabe [Boujelbane 2014]. En effet, la seule forme connue de ces dialectes est la forme orale familière est était absente dans tout document écrit. Récemment, les dialectes arabes ont acquis le statut des langues vivantes dans les études linguistiques cela a engendré l'émergence des efforts sérieux pour étudier les modalités et les régularités dans ces variétés linguistiques de ces dialectes. Cependant il faut signaler que ces derniers souffrent d'un manque de ressources et aussi l'absence des normes standards de transcription de ces dialectes. Bien que les dialectes arabes soient les porteurs de riches traditions orales pendant longtemps, ils ne sont pas des langues standardisés ou normalisés vu qu'ils ne sont pas enseignés dans les écoles.

C'est la raison pour laquelle il y en a de nombreux efforts qui s'imposent pour moderniser l'orthographe arabe et développer des orthographes pour ces dialectes. Ces efforts ont été en cours depuis de nombreuses années. Parmi ces travaux nous trouvons [Maamouri 2004] qui ont développé un ensemble de règles pour la transcription orthographique et l'annotation des dialectes du Levant, dans le but de créer un corpus levant arabe. De plus, [Habash 2012] ont proposé une convention de normalisation pour le dialecte egyptien (CODA). Cette convention vise à produire un équilibre optimal entre le maintien d'un niveau d'unicité dialectal et à établir des conventions sur la base de similitudes MSA-dialecte arabe. Leur convention a été utilisée dans leurs plusieurs outils pour les langages naturels et les ressources de traitement pour l'arabe Egyptien ([Habash 2013] ; [Eskander 2013]). Récemment, CODA a été également adaptée pour le dialecte algérien [Saadane 2015].

Afin dhomogénéiser les transcriptions orthographiques du le dialecte tunisien, il convient de créer des normes de transcription de ce dialecte. Ainsi, nous avons développé une variante CODA (Conventional Orthographique for Arabic Dialectal) pour le dialecte tunisien. En fait, notre travail est une continuation du travail de [Habash 2012] qui a proposé CODA, une orthographe conventionnelle pour l'arabe dialectal, qui est conçue dans le but de développer des modèles computationnels de dialectes arabes et a fourni une description détaillée de ses directives appliquées à l'arabe d'Egypte.

Ces objectifs rappelés, nous décidons dans la première partie de ce chapitre de présenter les ressources développées pour le traitement automatique du dialecte tu-

nisien. Nous donnons ensuite une vue globale des lignes directives de notre convention de normalisation CODA que nous avons développé. À cette fin, la section 4.2 expose les principes et les objectifs de CODA alors que la section 4.3 fournit une présentation détaillée pour les lignes directives de CODA. Dans la deuxième partie, nous présentons le recueil des corpus à savoir les signaux de parole et les transcriptions. En effet, nous décrivons brièvement la tâche d'enregistrement de notre corpus oral et la tâche de transcriptions. Enfin, nous présentons deux méthodes de collection de données textuelles essentiellement recueillies à partir des sites Web notamment les blogs Tunisiens et la translittération des données dialectales écrites en caractère latin.

4.2 Ressources développées pour le traitement automatique du dialecte tunisien

Rappelons que le dialecte tunisien possède des caractéristiques uniques qui le distingue des dialectes arabes ainsi que du MSA. Néanmoins, ce dialecte est considéré comme un langage under-ressources du fait qu'il ne possède ni orthographe standard, ni de grandes collections de textes écrits et de dictionnaires. Le progrès des recherches actuelles a contribué à la naissance de nouveaux corpus dialectal et de nouveaux outils et d'applications pour le traitement du dialecte tunisien. En loccurrence, [Graja et al., 2015] s'intéressent au problème de la compréhension littérale de la parole en dialecte tunisien dans le domaine de renseignement ferroviaire. À cet égard, ils ont proposées une méthode hybride qui consiste à combiner les modèles CRF et l'ontologie de domaine. Afin d'évaluer la méthode proposée, le corpus TuDiCoI (Tunisian Dialect Corpus Interlocutor) du dialogue oral en dialecte tunisien a été élaboré. Ce corpus est destiné pour la tâche de renseignements ferroviaires, et a été collecté en collaboration avec la Société Nationale des Chemins de Fer Tunisien. Ainsi, une convention orthographique baptisé OTTA pour le dialecte tunisien [Zribi et al., 2013] est proposé afin d'harmoniser les transcriptions. Nous signalons, entre autres, [Zribi 2013] se sont intéressés à l'analyse morphologique du dialecte tunisien. Comme cela, ils ont conçu de déterminer ses règles morphologiques et ses formes orthographiques pour les affixes, les clitiques, etc. Pour ce faire, ils ont exploités le corpus STAC (Spoken Tunisian Arabic Corpus). Ce corpus rassemble des transcriptions de quelques émissions télévisées et de la radio qui sintéressent à des domaines variés tels que la politique, la santé, les questions sociales et la religion. Dans le même ordre des idées, [Boujelbane 2014] ont mis l'accent sur la construction d'un corpus du dialecte tunisien afin de former un modèle de langue pour un SRAP. Ainsi, ils ont créé des ressources lexicales telles que le lexique bilingue MSA-dialecte tunisien. Pour ce faire, ils ont proposés une méthode de création de ressources à base de règles. Dans un autre exemple de travail, [Younes et al., 2015] ont construit un corpus contenant 43 222 messages

depuis des commentaires de Facebook, SMS et forums afin de s'en servir dans l'analyse automatique des messages dans ces réseaux.

4.3 Convention orthographique pour le dialecte tunisien (CODA)

CODA est une orthographe conventionnelle pour le MSA et tous les dialectes arabes, qui vise à combler la lacune d'absence de norme d'écriture des dialectes arabe conçue principalement pour la proposition d'élaborer des modèles informatiques pour les dialectes arabes. De ce fait, CODA aborde trois orthographes enjeux : ***i)*** Fautes de frappes simples telles que des lettres transposées ou des espaces manquants ou fautes d'orthographe, ***ii)*** Des mots orthographiés créatifs intentionnellement et les effets de la parole. ***iii)*** Un choix de l'orthographe dialectal inconsistant cause le manque des normes.

4.3.1 Les objectifs de CODA

Généralement, [Habash 2012] ont présenté CODA selon cinq objectifs :

Conventional Orthographique for Arabic Dialectal CODA	
Objectifs	1. CODA est une convention de cohérence interne pour l'écriture arabe dialectale (DA). Théoriquement dans CODA, chaque mot a une seule représentation orthographique.
	2. CODA est créé pour besoins computationnelles.
	3. CODA utilise l'alphabet arabe.
	4. CODA est conçu comme un cadre unifié pour l'écriture de tous les dialectes arabes.
	5. CODA vise à trouver un équilibre optimal entre le maintien d'un niveau d'unicité dialectale et d'établir des conventions sur la base de similitudes MSA-DA.

Figure 4.1 – Les objectifs de CODA.

4.3.2 Les principes de CODA

CODA est une orthographe conventionnelle pour l'arabe dialectal. Aussi il faut mentionner que sa conception respecte plusieurs principes :

- **Une convention ad hoc** : premièrement, CODA est une convention ad hoc. Il existe de nombreuses décisions qui pourraient avoir des effets différents surtout quand il s'agit de l'interface orthographe-phonologie.
- **Écriture Arabe** : deuxièmement, CODA utilisent uniquement les caractères arabes, y compris les signes diacritiques pour écrire dialectes arabes. Cependant, CODA interdit l'utilisation des caractères arabes étendus, par exemple, celle de la langue persienne. Tout comme le MSA, CODA permet d'écrire sans diacritiques.
- **Cohérence** : troisièmement, CODA conserve la cohérence des mots. En d'autre termes, chaque mot a une forme orthographique unique dans CODA qui représente sa phonologie et sa morphologie.
- **Ressemblance au MSA** : encore, en profitant de ressemblances entre le MSA et le dialecte arabe, CODA emploie les décisions orthographiques (règles, les exceptions et les choix ad hoc).
- **Phonologie** : de même, CODA conserve généralement la forme phonologique des mots dialectaux étant donné les règles phonologiques de chaque dialecte, et les limites de l'écriture arabe.
- **Morphologie et syntaxe** : de plus, elle préserve la morphologie dialectale et la syntaxe dialectale.
- **Lisibilité** : en règle générale, CODA est facile à lire et à apprendre. Ainsi, l'ensemble des dialectes arabes possèdent généralement les mêmes principes de CODA, où chaque dialecte disposera de son unique correspondance CODA qui respecte sa phonologie et sa morphologie. Toutefois, CODA n'est pas une représentation purement phonologique.

Dans ce qui suit, nous allons présenter nos lignes directives de CODA.

4.4 Les Lignes directives de CODA pour le dialecte tunisien

Nous présentons par la suite des lignes directives spécifiques pour CODA du dialecte tunisien conformément aux mêmes règles orthographiques que MSA avec les exceptions et les extensions bien déterminées.

4.4.1 Les extensions phonologiques

4.4.1.1 Système vocalique

En dialecte tunisien, nous relevons la distinction dans la prononciation de certaines voyelles longues. Comme mentionné dans l'exemple du mot حرَام /HrAm/

[religieuse interdite] où la voyelle longue fatha suivie alif se prononce de la même manière en dialecte tunisien et en MSA. Tandis que cette voyelle longue se prononce comme أي /Ay/ en dialecte dans l'exemple de mot حرَام /HrAim/ [un vêtement traditionnel]. Ainsi, pour faire la distinction entre ces deux prononciations possibles de cette voyelle longue nous avons proposés d'ajouter une autre voyelle longue inexistant en MSA qui est kasra suivie par Alif afin d'exprimer le phonème de أي /Ay/. Par voie de conséquence, le dialecte tunisien possède une longue voyelle kasra suivie par Alif qui n'existe pas en MSA.

Dans d'autres cas, nous avons remarqués que lors de la prononciation de certains mots dialectal un phénomène de raccourcissement de voyelles longues ou des clitiques est apparu. Prenons à titre d'exemple le mot تقول /tqwl/ [dit] lorsque il est suivi par l'un de ces deux clitiques : لهَا /lhA/ ou له /lh/ on remarque un raccourcissement de ces clitiques de plus une combinaison des deux lettres ل /l/ comme s'ils ont porté le shadda. Autrement dit, le clitique لهَا /lhA/ devient هَا /hA/ et le clitique له /lh/ devient لو /lw/. Ainsi, nous avons profité des ressemblances entre le MSA et le dialecte tunisien pour prendre une décision d'écrire orthographiquement ces deux mots comme en MSA تقول لهَا /tqwl lhA/ [tu lui dis] et تقول له /tqwl lh/ تقول له /tqwl lh/.

4.4.1.2 Système consonantique

- **Le cas de « hamza »**

Nous avons déjà indiqué auparavant que le graphème « hamza » possède différentes formes. En fait, en dialecte tunisien ce graphème peut subir plusieurs modifications lors de la prononciation. De ce fait, nous distinguons trois types de « hamza » que nous présenterons par la suite.

- Premièrement, nous trouvons le « hamza » réel qu'on appelle en arabe همزة قطع /hamzet qatE/ qui est toujours prononcé en tant qu'un «coup de glotte» situé soit au début ou au milieu d'un mot. Par exemple إمَام /AmAm/.
- Deuxièmement, la temporaire « hamza » همزة وصل /hmzt wasel/ représente le son de « hamza » associé à la lettre Alif ajoutée pour les mots commençant par une voyelle. En effet, ce « hamza » ne sera pas conservée si le mot est un verbe à la voix passive. Dans ce cas, on écrit les mots orthographiquement sans « hamza » comme le mot تكسر /tksr/ [brusé] malgré sa forte prononciation avec «hamza» إتكسر /Atksr/.

- Le troisième type est celui du « hamza » de MSA qui est disparait ou remplcé par autre en dialecte tunisien. Autrement dit, il est présenté dans la partie graphèmique et phonémique des mots en MSA, cependant, il disparait ou remplacé par d'autre graphème lors de leurs prononciations en dialecte tunisien.

En règle générale, nous distinguons trois façons d'écrire la lettre « hamza » qu'on va les citer avec des exemples par la suite :

Tout d'abord, la première façon traite le « hamza » situé au milieu d'un mot. Ce « hamza » peut être à l'orale du dialecte tunisien silencieux ou remplacé par le son de « ya ». Prenons l'exemple de mot الدقّاءق /AldaqAiq/ [les minutes] où le « hamza » du MSA est remplacé par le son de « ya ». Ainsi, ce mot est prononcé en dialecte en tant que الدقّايق /Aldaqayaq/. Dans le même ordre des idées, lorsque le « hamza » situé au milieu d'un mot comme dans l'exemple de mot dialectal فءرَان /fArAn/ [les souris], dans ce cas le « hamza » du MSA devient silencieux et le mot se prononce de cette façon فيرَان /fyrAn/. Cependant, nous avons signalés qu'il y a des exceptions où le « hamza » est prononcé en MSA et en dialecte tunisien de la même façon comme dans l'exemple de mot أسءلة /sAlAt/ [des questions]. Ainsi nous avons décidé d'écrire orthographiquement le « hamza » en dialecte comme ça se prononce.

La deuxième façon s'intéresse au « hamza » qui s'écrit au milieu ou à la fin d'un mot mais en dialecte tunisien elle devient « alif » comme dans l'exemple كأس, nous remarquons que le « hamza » se prononce en MSA par contre en dialecte tunisien le mot se prononce comme كاس. Cela veut dire, le « hamza » de MSA devient « alif » en dialecte tunisien. De même, dans la deuxième façon il y a des exceptions tel que le mot سأل /sAl/ [interroge] qui se prononce de la même façon en MSA et en dialecte. D'où nous avons décidé d'écrire le « hamza » comme ça se prononce en dialecte.

La troisième façon concerne le « hamza » qui s'écrit en MSA au milieu d'un mot mais en dialecte tunisien devient « waw» comme dans l'exemple du mot رأس /rAws/ [les têtes] d'où ce mot se prononce comme روس. Semblable à la première et la deuxième façon, il y a des mots d'exception où le « hamza » se prononce de la même façon en MSA et en dialecte comme le mot مسُوّلية /msAwlyap/ [responsabilité]. Par conséquent, nous avons décidé d'écrire le « hamza » comme ça se prononce en dialecte tunisien.

- **le cas des autres consonnes**

Au surplus, il s'agit d'une autre extension phonétique liée au système consonantique du dialecte tunisien. C'est le cas de la double prononciation de la lettre ق /q/ et la variation de prononciation de certaines consonnes. En somme, nous avons pris la décision d'écrire l'ensemble de consonnes sous une forme qui reflète l'apparenté à la racine MSA. De ce fait, nous distinguons dans CODA deux cas spécifiques :

Comme déjà signalé dans le premier chapitre, en dialecte tunisien la consonne ق /q/ possède une double prononciation ق /q/ et ڤ /G/. Ainsi, selon la décision de CODA, la lettre ق /q/ est utilisée pour représenter les deux consonnes /q/ et /G/. Cependant, il existe quelques exceptions où le ق /q/ toujours se prononce comme ڤ /G/, donc selon CODA cette consonne r. comme : قَازوز /boisson gazeuse/ ou منقَالة [minge :la] /une montre/.

Nous avons mentionné qu'en dialecte tunisien nous pouvons trouver quelques variations dans la prononciation de certaines consonnes. Étant donné que parmi nos objectifs de la création de CODA est d'avoir pour chaque mot une seule représentation orthographique, nous avons décidé d'écrire la consonne telle qu'elle figure en MSA dans le cas où il y a des variations de prononciation. À titre d'exemple, la consonne س /s/ qui peut être prononcée comme س /s/ ou ص /S/. Partant de ce fait, le mot رسول /rswl/ [prophet] peut être prononcé رسول ou رصول. Cependant, ce mot existe en MSA et s'écrit رسول /rswl/ avec la consonne س /s/, alors, nous devons l'écrire en dialecte tunisien comme رسول /rswl/ en gardant la même forme de MSA.

4.4.2 Les extensions morphologiques

En règle générale, la langue arabe est fortement agglutinante comprenant des articles, des conjonctions, des prépositions, matérialisées par des clitiques qui se rattachent aux formes fléchies. Généralement, les proclitiques qui se situent avant la forme fléchie et les enclitiques qui se situent après sont distingués. De surcroît, le mot de base se compose d'une racine et affixes obligatoires (dont certains sont nuls). Cependant, pour le mot de base, des clitiques et des proclitiques peuvent être ajoutés d'une façon optionnelle. Le tableau 4.1 présente un exemple illustratif des

enclitiques et proclitiques du mot du dialecte tunisien و شريتوهَاشي /w$ritwhA$y/ (L'avez-vous acheté ?).

Enclitiques	Enclitiques	Suffixes	la base des tiges	proclitiques
شي	هَا	و	شريت	و

Table 4.1 – Un exemple illustratif des enclitiques et proclitiques du mot و شريتوهَاشي /w$ritwhA$y/ (L'avez-vous acheté ?).

Ce qui suit présente une liste des affixes et des clitiques spécifiques au dialecte tunisien.

4.4.2.1 Les affixes

Les noms et les adjectifs prévoient des règles spécifiques pour les affixes :

- Ta-marbouta est toujours ة et ne pas être ه à la fin du mot. Dans l'exemple كرهبة سميرة جديدة/krhbt samyrap jdydap/ [la voiture de Samira "nom propre" est neuve] bien que à la fin des deux mots سميرة /samyrap/ [nom propre] et جديدة /jdydap/ [neuve] le ta-marbouta est silencieux, on l'écrit comme dans le MSA ة /t/.
- Les suffixes nominaux du dialecte tunisien sont ات /At/ et ين /yn/ qui désigne respectivement le double et le pluriel. D'où, nous avons gardé les mêmes suffixes nominaux de MSA.
- Les affixes pluriels وَا /wA/ et توَا /twA/ sont prononcés avec un « alif » silencieux à la fin du mot qui est le même cas pour le MSA tels que dans les deux cas : كتبوَا /ktbwA/ [ils ont rédigé] et كتبتوَا /ktbtwA/.

4.4.2.2 Les clitiques

Nous avons distingué deux types de clitiques à savoir, les clitiques attachés et les clitiques séparés.

- **Les clitiques attachés** : Outre les clitiques et les proclitiques de MSA qui sont l'article défini /Alif et Lam/, la conjonction de coordination و /w/ [et], les articles ل ،ك ،ف ،ب /b, f, k,l/, les pronoms de la 3ème personne

ه ، هَا، هم /h, hA, hm/ et les enclitiques pronominaux نَا، كم ، هم، هَا /nA, km, hm, hA/, nous trouvons celles du dialecte tunisien qui sont : le proclitique d'interrogation شي /$y/ et l'enclitique de négation des particules ش /$/. En effet, la particule d'interrogation شي /$y/ est accordée aux verbes imparfaits ou aux des verbes parfaits. Par exemple : dans le verbe parfait عملت /Emlt/ [fait], lorsqu'on lui ajoute le proclitique d'interrogation شي /$y/, il devient عملتشي /Emlt$y/ [tu fait]. Par ailleurs, dans le verbe imparfait تعمل /tEml/ lorsqu'on lui ajoute le proclitique d'interrogation شي /$y/ il devient تعملشي /tEml$y/.

- **Les clitiques séparés** : La forme négative d'un verbe sans la particule de négation n'a pas de sens en dialecte tunisien, par exemple كتبش /ktb$/. Ainsi, selon CODA tunisienne il faut préserver la règle des enclitiques d'objets indirects et également le proclitique de négation qui exige la séparation avec un espace entre la particule de négation et les enclitiques d'objets indirects, par exemple مَا + قَال + ليش /mA + qAl + ly$ [il ne m'a pas dit].

4.4.3 Les exceptions lexicales

CODA du dialecte tunisien comporte une liste de mots spécifiant l'orthographe des mots tunisiens qui ont une orthographe exceptionnelle ou qui sont couramment orthographiés de différentes manières et nécessitent clairement le choix de CODA. Ainsi, notre objectif consiste à avoir un corpus cohérent où chaque mot possède une seule représentation orthographique comme dans le cas des pronoms أنتي /Anty/ [tu en féminin] qui s'écrit de cette manière et non pas comme أنت /Ant/ [tu]. En d'autres termes pour le pronom démonstratif هذَاكة /h*Akt/ on l'écrit de cette manière et non pas comme هَاذَاكة /hA*Akt/. De plus, de nombreux mots étrangers sont utilisés et même intégrés dans le dialecte tunisien. Ces mots sont écrits dans notre texte avec des caractères arabes.

Tout au long de la première partie nous avons défini notre convention de normalisation pour le dialecte tunisien CODA tout en mettant l'accent sur ses principes et ses objectifs. De surcroît, nous avons précisé les lignes directives de CODA. Compte tenu de ces lignes directives, nous allons recueillir nos ressources dialectales qui vont être utilisées pour le développement d'un SRAP pour le dialecte tunisien. Ainsi, les sections suivantes décrit notre méthode pour la construction des ressources.

4.5 Corpus de renseignement ferroviaire Tunisien

La collecte rapide et facile d'une grande quantité de corpus est une tâche essentielle pour le développement d'un SRAP contenant un vocabulaire abondant dans une nouvelle langue. Ces données recueillies comportent d'une part des signaux de parole pour l'apprentissage des modèles acoustiques du système, et d'autre part des données textuelles pour l'apprentissage des modèles statistiques du langage du système. Néanmoins, ce type de ressource n'est pas disponible pour certaines langues connues comme des « langues peu dotées ». Le dialecte tunisien appartient à cette catégorie de langues peu dotées. Ainsi, ce dialecte se plaint d'un manque de ressources et d'outils indispensable pour son traitement automatique. De ce fait, ce défi constitue notre principale motivation pour la conception et la réalisation d'un SRAP pour le dialecte tunisien.

Dans ce contexte, nous avons aborder le problème du manque de ressources du dialecte tunisien en développant une méthodologie de collecte de ressources : corpus de parole avec le texte correspondant.

Dans la deuxième partie, nous présentons tout d'abord le recueil des signaux de parole. Nous décrivons ainsi brièvement la tâche d'enregistrement de notre corpus oral, en indiquant toutes les dispositions prises en compte, les conditions ainsi que le lieu d'enregistrement. Cela va de soi, ces enregistrements sont prêts à être transcrits. À ce propos, nous traitons la tâche de transcriptions et les outils utilisés pour réaliser la récupération d'un corpus de texte.

4.5.1 Enregistrement

Notre objectif principal dans ce travail est de développer un SRAP dans le domaine du renseignement ferroviaire Tunisien. Afin de surmonter le problème d'absence des enregistrements audio de bonne qualité répondant à nos exigences, nous avons décidé de construire notre propre corpus basé sur des enregistrements réels.

Après avoir eu l'autorisation des personnels de la gare de Tunis, nous avons pu enregistré dans des conditions réelles. En outre, le choix de cette gare est justifié par le fait que Tunis est la capitale de la Tunisie qui réunit les diverses catégories des citoyens venant de différentes régions Tunisiennes. Ce qui implique l'existence de différents accents. De surplus, la variété du niveau culturel et social donne une richesse de vocabulaire que nous ne pouvons pas trouver dans une autre région de la Tunisie.

En effet, les expériences ont été faites dans deux guichets de la gare et à chaque fois les enregistrements sont été effectués par un agent volontaire. De cette façon,

nous avons enregistré les conversations entre les clients et les agents dans lesquelles il y avait une demande d'informations concernant les horaires du train, les tarifs, les réservations, etc. En ce qui concerne les équipements utilisés, nous nous sommes servis de deux PC portables en exécutant le logiciel Audacity et deux micros, l'un pour l'agent de guichet et un autre pour le client. Il faut mentionner que le micro du client était invisible. Cela nous a permis déviter toute perturbation et hésitation de la part du client et puisse parler d'une manière spontanée. En outre, ces conversations sont faites dans de différentes périodes telles que les périodes des vacances, les week-ends, les jours de fête, et parfois au cours de la semaine. Delà, notre objectif a été d'avoir un corpus varié le maximum avec un vocabulaire riche, étant donné que les tarifs et les horaires changent dans les périodes de vacances par rapport au cours de l'année en plus les jours de promotion et les jours de fêtes.

Tout compte fait, nous avons obtenu 20 heures d'audio après une période de 3 mois d'enregistrement. En fait, la thématique principale de ce corpus d'audio est la demande d'informations sur les services de chemins de fer dans une gare en dialecte tunisien. Ces demandes correspondent aux types de train, ses horaires, sa destination, le prix et la réservation des billets.

La mise à disposition d'un tel corpus se heurte en premier lieu aux aspects juridiques notamment le respect de la vie privée. Nous verrons que l'anticipation de ceux-ci a eu une incidence sur l'ensemble de la chaîne de traitement.

4.5.2 Respect de la vie privée

Comme le recueil de notre corpus de parole est basé sur des conversations réelles, nous pouvons y trouver des données personnelles entre les agents et les clients. Il s'agit des formes nominatives, des professions, statuts, ou titres, des activités sociales, des liens de parenté, des réseaux, des références à des lieux et/ou des références à des caractéristiques de la personne comme les numéros des cartes d'identité nationale « CIN », les numéros de carte de travail, les numéros de carte de fidélité... Ainsi, la présence de ces données personnelles dans un corpus engendre un manque de respect de la vie privée des gens et implique aussi une mise en conformité avec la loi informatique et liberté. En suivant le Guide des bonnes pratiques 2006 pour manœuvrer les droits liés au respect de la vie privée, il faut surveiller méticuleusement le cadre légal de gestion des données personnelles des gens et de procéder à l'anonymisation à ces données. Juridiquement, l'anonymisation ne permet pas de découvrir l'identité d'une personne. De plus, nous avons établi une convention avec l'administration de la SNCFT pour éliminer les indices permettant d'identifier directement la personne, ainsi que les éléments qui peuvent lui porter préjudice.

En fait, notre idée de base consiste premièrement à repérer les éléments sensibles existant dans les enregistrements et qui peuvent donner une information personnelle sur le locuteur « client ». D'autre part, nous avons insisté sur l'anonymisation des données personnelles en supprimant les indices qui permettent d'identifier directement le client.

En observant le corpus, nous avons constaté que c'est souvent le regroupement de plusieurs indices qui permet le dévoilement de l'identité du locuteur. Par exemple, un voyageur donne son prénom sans indiquer son nom, cette information n'est pas considérée en tant qu'une identification de la personne. De plus, être un journaliste ne permet pas d'identification, en indiquant le nom de son journal par exemple « journal la presse » ne présente pas une identification, sauf s'il précise son nom, dans ce cas on peut arriver à un singleton. À cela s'ajoutent les exemples comme «Je suis le président du centre police de la circulation du zone de Bardo» qui sont rarement présents dans le corpus et permettent, en revanche, l'identification directe du locuteur.

En général, l'anonymisation s'effectue manuellement en vérifiant la totalité du dialogue déroulé entre un agent et un client tout en supprimant des données qui identifient le client. Il vaut mieux précéder la procédure d'anonymisation en vue de simplifier le processus de transcription.

En achevant les enregistrements nous pouvons passer à la transcription afin d'obtenir un corpus de texte. Les sections suivantes sont consacrées à la description des opérations de traitement du corpus entrant en jeu dans la préparation de ressources destinées d'une part à la réalisation d'un SRAP et d'autre part à une libre diffusion.

4.5.3 Outil d'aide à la transcription orthographique : Transcription

La transcription qui est le premier degré d'une représentation symbolique de l'oral est une étape primordiale dans la constitution du corpus. La qualité de ces transcriptions faites à ce stade influence donc tout le traitement postérieur. Cependant, la transcription en elle-même n'est pas une tâche anodine. La tâche a été d'autant plus difficile qu'il n'y a pas de conventions de transcription pour le dialecte tunisien.

Plusieurs contraintes ont influencé nos choix. Notre objectif principal était d'avoir un corpus audio et des données transcrites de bonne qualité et suit les normes de notre convention de normalisation CODA. Le processus de transcription devait donc être effectué rapidement mais avec une bonne efficacité. Étant donné la non-existence d'outils de transcription automatique disponibles pour le

dialecte tunisien, le recours à la transcription manuelle est crucial. De ce fait, notre première phase consiste à choisir trois étudiants de l'université pour réaliser cette tâche. Certes, il s'agit de la transcription orthographique où chaque conversation que nous pouvons appeler « dialogue » entre un agent et un client est gardée séparément. Autrement dit, il existe un tour de parole qui correspond à l'occupation matérielle du canal de parole par un locuteur, ce tour s'achève lorsqu'un nouveau locuteur prend la parole à son tour. Dans nos transcriptions, la succession des tours de parole est présentée d'une façon horizontale : les paroles des locuteurs se succèdent, de haut en bas et chaque paragraphe présente l'intervention d'un locuteur « client ».

De ce fait, dans notre corpus nous avons gardé chaque dialogue entre un client et un agent. Cette conversation est composée de plusieurs demandes susceptibles d'être combinées ensemble au cours d'un dialogue entre le personnel et le client au sujet des services ferroviaires à la gare. Un exemple d'un véritable dialogue en dialecte tunisien entre un client et un agent est illustré dans la figure 4.2. Selon cet exemple, deux types de locuteurs ont participé à ce dialogue qui sont un client et un agent.

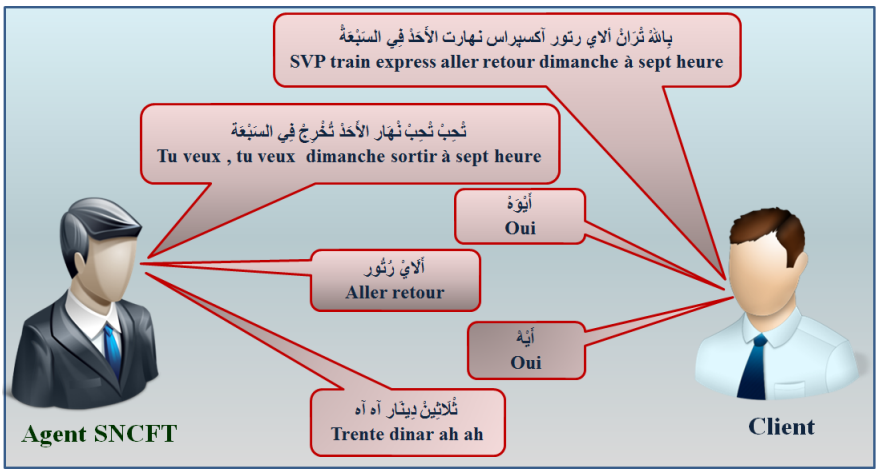

Figure 4.2 – Exemple d'un dialogue réel en dialecte tunisien entre un client et un agent.

4.5.3.1 Transcriber

L'alignement de ces transcriptions avec le son était une autre contrainte. Ainsi, nous devons naviguer dans la transcription et le son en parallèle. L'objectif a été défini de transcrire et rendre disponible l'intégralité du corpus en y associant des intervalles temporels. Dans la littérature il existe plusieurs logiciels utilisés pour réaliser la transcription orthographique et l'annotation d'un fichier audio comme

Transcriber [1], Praat [2], ANVIL [3] , ELAN [4]...

Notre choix s'est arrêté sur le logiciel de transcription Transcriber qui répondait complètement à nos attentes et qui permet de réaliser des transcriptions de l'oral alignées avec le signal. Les raisons de notre choix tiennent à la facilité d'usage de ce logiciel au moment de la transcription et au moment de son utilisation. À l'aide de ce logiciel, nous pouvons récupérer des fichiers ayant une extension .trs dans le format original produit par Transcriber permettant de coder la synchronisation du son et de sa transcription. Le corpus a été découpé au niveau du tour de parole (voir Annexe A).

4.5.3.2 Conventions de transcription avec Transcriber

Toute transcription est un compromis forcément boiteux entre le respect des particularités orales et la lisibilité. Afin d'imposer la lisibilité, nous avons adopté une transcription orthographique pour faciliter la lecture cursive. Tout au long de la transcription de notre corpus à l'aide de l'outil Transcriber, nous avons suivi les normes décrites dans notre convention de normalisation CODA. Outre ces normes de CODA, nous avons produit des normes pour la transcription à l'aide de Transcriber pour mettre les données plus lisible.

Afin de rendre notre corpus disponible et exploitable, nous avons suivi les principes d'annotation suivants pour la transcription : lisibilité, conservation des spécificités de l'oral, volonté d'un maximum d'interopérabilité et codage non ambigu.

Dans les sous sections suivantes, nous allons décrire les normes que nous avons fixées pour la transcription à l'aide du Transcriber.

- **Voyellation de corpus**

De toute évidence, la norme orthographique de notre convention CODA ne pose aucune obligation sur la voyellation des textes puisque chacun a le droit de choisir en fonction de ses exigences d'avoir des transcriptions voyellées ou non. Généralement, dans le cas où les textes écrits en dialecte tunisien sont non voyellés, nous sommes face à une ambiguïté dans le traitement automatique étant donné que la même forme d'un mot sans voyelles peut correspondre à plusieurs mots voyellés. Malgré qu'oralement toutes les voyelles sont présentes et bien prononcées.

1. http ://trans.sourceforge.net/en/presentation.php.
2. http ://www.fon.hum.uva.nl/praat/
3. http ://www.anvil-software.de/
4. http ://icar.univ-lyon2.fr/projets/corinte/confection/elan.htm

Ainsi, dans le but d'enlever cette ambiguïté et de pallier ce manque de voyelles, beaucoup de travaux de recherches ont été proposés dans la communauté du TALN tels que les travaux de voyellation automatique de textes ([Ryan 2008] ; [Elshafei 2006]. Ces travaux sont dédiés principalement pour le MSA. Par contre, jusqu'à maintenant, il n'existe aucun travail sur la voyellation automatique du dialecte tunisien. Par voie de conséquence, nous sommes mis d'accord de voyeller les mots lors de la transcription manuelle de notre corpus pour avoir des conversations plus précises et non ambigües.

Selon l'étude de notre corpus, nous avons rendu compte qu'il existe quatre types de mots : les mots en MSA avec voyellation dialectal ; les mots en dialecte (DT) ; les mots en MSA contenant des affixes en dialecte (DT*), des mots étrangers (ME) et enfin des mots étrangers qu'on utilise avec des modifications par rapport à leurs origines (ME*). En somme, (ME*) est un mot étranger qui a subi un ajout d'enclitiques ou de proclitiques de la langue arabe.

Dans ce qui suit, nous allons présenter dans les deux figures 4.3 et 4.4 un extrait de notre corpus pour montrer les quatre types de ces mots.

Phrase en Dialecte Tunisien		سُوسَة	ترينو	خَرَجْ	عَسْلامَة
Traduction en Français		Sousse	Train	Parti	Salut
L'origine des mots	MSA	X		X	
	ME		X		
	ME*				
	DT				X
	DT*				

Figure 4.3 – Exemple 1 de mots en dialecte tunisien.

Phrase en Dialecte Tunisien		ريزرفيلي	الأوْقَاتْ	مَا نعْرَفْشْ	أنَا
Traduction en Français		réserver moi	les horaires	ne sais pas	je
L'origine des mots	MSA		X		X
	ME				
	ME*	X			
	DT				
	DT*			X	

Figure 4.4 – Exemple 2 de mots en dialecte tunisien.

Selon ces deux exemples, le mot ترينو [Trinou] est un mot étranger d'origine italienne. Le mot مَا نعرفش /ne sais pas/ est un mot en MSA contenant des affixes dialectal مَا /ne/ et ش /pas/. Ces deux affixes expriment la négation en dialecte

tunisien. Enfin, Le mot ريزرڤيلي /réserver moi/ est un mot emprunté du français, il subit un ajout de l'enclitique arabe لي /moi/ qui est attaché au mot.

- **Les mots étrangers**

À priori, le dialecte tunisien se caractérise par la présence des mots empruntés en français, en berbère, en italien, en Turc et en espagnol. De plus, l'existence des mots dont l'origine est MSA. En fait, les normes de CODA imposent seulement l'écriture des mots avec l'alphabet arabe ayant comme origine MSA ou l'arabe dialectal. Toutefois, il n'existe aucune contrainte sur l'écriture des mots étrangers comme le français et l'anglais. Par conséquence, nous avons choisi d'écrire ces mots avec l'alphabet arabe. Pour transcrire ces mots, l'alphabet arabe a été utilisé, néanmoins, il y a des lettres latines comme G, V, P qui ne correspondent pas à des lettres arabes qui sont donc définis comme suit : /G, ڤ/, /V, ڤ/ et /P,پ/.

Ainsi, chaque mot étranger possède une forme unique par exemple : /Première, پرميَار/, /Retour, رتور/ et /Express, أكسپرَاس/.

- **Valorisation de l'oralité des corpus**

Notre corpus est constitué de conversations réelles, autrement dit, il s'agit de la parole spontanée. Ainsi, les «disfluences» représentent un phénomène fréquent dans la production orale spontanée. Par ailleurs, il faut indiquer que les principaux phénomènes de ces «disfluences» [Bahou 2010] sont les répétitions, les autocorrections, des hésitations et des mots incomplets représentent un phénomène qui se produit fréquemment dans l'oral spontané. Durant la transcription de notre corpus, nous conservons les traces de ces phénomènes de « disfluences ».

— **Les répétitions** : consistent à répéter un mot ou une suite de mots. Théoriquement, la plupart des répétitions dans notre corpus sont utilisées par un locuteur pour affirmer ou reformuler sa demande. La figure 4.5 représente un exemple de répétition dont laquelle le locuteur affirme sa demande.

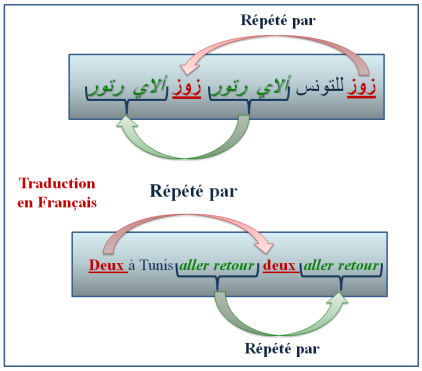

Figure 4.5 – Répétition dont laquelle le locuteur affirme sa demande.

Dans l'exemple 4.6, la répétition est effectuée par le locuteur pour présenter sa demande. Il a utilisé deux mots différents ayant la même sémantique.

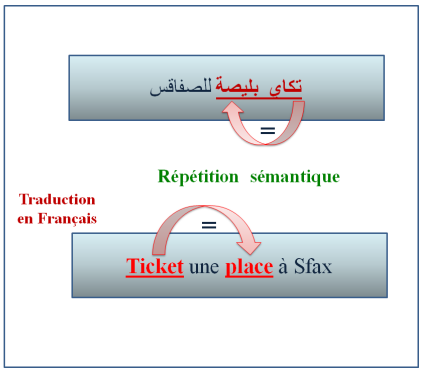

Figure 4.6 – Exemple de répétition sémantique, extrait de notre corpus TARIC.

— **Les auto-corrections** : le locuteur peut commettre une ou plusieurs erreurs et de les corriger dans le même énoncé. Ce phénomène ressemble beaucoup à une répétition, mais la partie répétée est une reconstruction d'une mauvaise partie dans l'énoncé. La figure 4.7 présente un exemple d'auto-corrections.

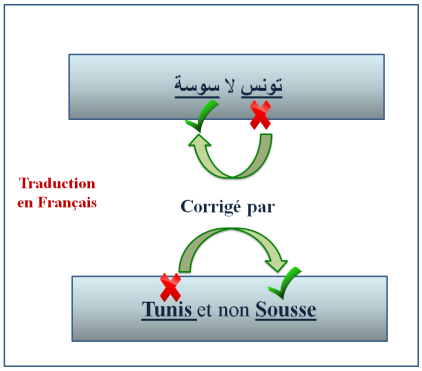

Figure 4.7 – Exemple d'auto-corrections.

— **Les hésitations** : ce sont des phénomènes qui apparaissent dans l'expression orale spontanée. Elles peuvent se manifester de diverses façons : soit en utilisant un morphème spécifique (par exemple, ا ا, اه, etc.) ou sous forme d'un allongement de la syllabe. Il s'agit de classes lexicales faisant partie seulement à la production orale spontanée.

L'exemple dans la figure 4.8 illustre un exemple de marqueur d'hésitation présent dans notre corpus.

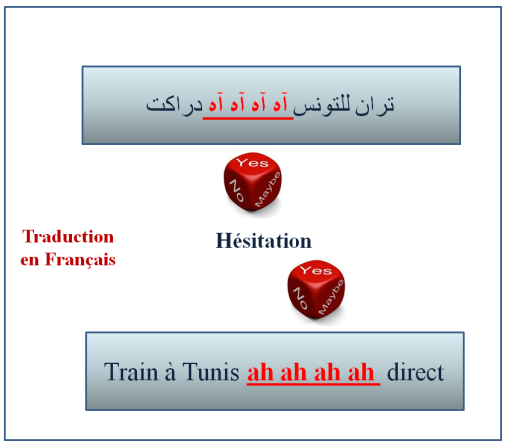

Figure 4.8 – Un exemple d'hésitation.

— **Les amorces** (**mot incomplet**) : ce sont les cas d'arrêt de la production d'un mot avant la fin normale de celui-ci. Dans sa terminologie, un mot incomplet est toujours un fragment de texte qui peut être identifié par la connaissance de la phraséologie. L'amorce d'un mot est notée par un tiret accolé au mot :

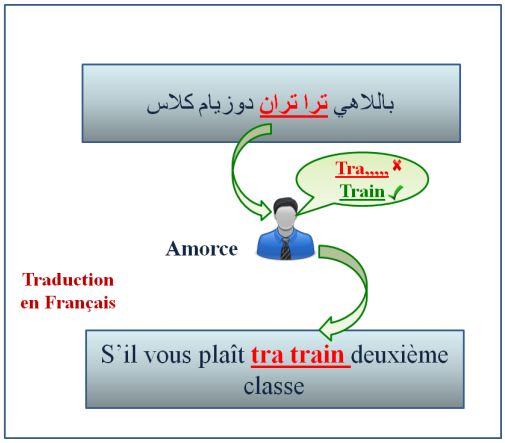

Figure 4.9 – Exemple d'amorce extrait de notre corpus TARIC.

Dans l'exemple de la figure 4.9, le locuteur tente à prononcer le mot «train» mais il s'arrête avant la fin normale du mot. Puis il commence à prononcer à nouveau le mot complet.

Tout bien considéré, ce travail demande une grande attention de la part du transcripteur, pour noter ces phénomènes qui sont communément flou dans une écoute ordinaire.

- **Chevauchement de parole**

Dans les conversations, il est très fréquent que deux locuteurs (agent et client) parlent en même temps. Le chevauchement de la parole en elle-même pose de toute façon à un problème : il faut savoir à distinguer les différents locuteurs, ainsi que leurs propos respectifs. Le logiciel Transcriber est mal adapté à la notation des chevauchements, en particulier il ne permet pas de noter le cas où plus de deux personnes prennent la parole simultanément. Dans notre cas, nous avons utilisé d'autres artifices (bruit de fond, parole superposée) pour préciser qu'il y a un chevauchement de parole.

- **Ponctuations**

La ponctuation permet d'organiser et de présenter la transcription, d'une part, et de faciliter la compréhension de la transcription, d'autre part. La question qui se pose est selon quels principes on va mettre un point, une virgule et un point-virgule ou bien les éliminer. Par ailleurs, le point d'interrogation, le point d'exclamation et les guillemets ont été utilisés lorsque le transcripteur entendait nettement l'intonation : ? Interrogations avec montée de la voix ! Exclamation

En clôturant ces transcriptions, nous avons obtenu un corpus composé de 20 heures d'audio avec la présence de sa transcription orthographique. De ce fait, pour chaque audio, il lui correspond un fichier Transcriber. Ce corpus est nommé TARIC : Tunisian Arabic Railway Interaction Corpus. Notre corpus TARIC se compose de 4662 dialogues et représente 18657 d'énoncés nous allons par la suite présenter les caractéristiques les plus importantes dans le tableau 4.2.

Caractéristiques	***Les chiffres***
Dialogues	4662
Nombre d'heures	20
Nombre d'énoncés	18657
La taille du vocabulaire	71684

Table 4.2 – Les caractéristiques principales du corpus TARIC.

Ce corpus TARIC que nous avons recueilli a atteint 4000 mots non redondantes. Celle-ci est considérée comme une quantité limitée pour la conversion G2P. De ce fait et dans le but de pallier le problème de la carence des données, nous avons essentiellement recueilli des ressources textuelles issues à partir de deux méthodes de collection de données : Aspiration des blogs et Translittération des données dialectales écrites en caractère latin.

Nous abordons dans les deux sections suivantes ces deux méthodes de recueil de ressources pour le dialecte tunisien.

4.6 Aspiration des blogs

Depuis l'émergence de l'internet, la puissance informatique a été développée et a permis une grande facilité de stockage et les données de texte ne cessent de croître. De surcroît, le Web et les journaux électroniques ont facilité la collecte des textes d'une façon gratuite, rapide et avec une quantité de textes satisfaisante accessible pour de nombreuses langues. De ce fait, afin de collecter des corpus textuels en grande quantité, plusieurs chercheurs ont recours à l'Internet en se basant sur une approche intéressante qui consiste à « aspirer » un grand nombre de sites Web dans la langue donnée et à filtrer les données récupérées pour les rendre exploitables. Néanmoins, au niveau des langues peu dotées le recours à cette méthode rencontre de nombreux problèmes concernant le nombre limité des sites Web, la faible vitesse de transmission et la qualité variable des documents ce qui demande alors plus d'outils de traitements. Dans ce cadre, notre problématique se résume en la quasi-absence des sites web écrits en dialecte tunisien vu que la plupart de ces sites sont présentés en MSA ou en Français à l'exception des blogs.

Généralement, le Blog est nommé par la contraction des mots Web et Log (carnet de bord web en anglais). C'est un type de site web utilisé pour la publication périodique et régulière d'articles et rendant compte d'une actualité autour d'un sujet donné. Quand nous parlons des blogs il faut mentionner qu'il existe de multiples formes de blogs qui traitent des thèmes variés parmi lesquels : le journal intime, le blog de journaliste, le blog d'hommes politiques aussi le blog d'actualité, etc. En outre, les blogs sont aussi une nouvelle source d'informations qui n'existaient pas avant et qui permettent aux internautes de se communiquer entre eux par le biais de discussion sur tous les sujets.

Ainsi, dans le but de solliciter la problématique du manque de ressources de cette langue, nous avons choisi de collecter les données à partir des sites de blogs écrits en dialecte tunisien (écrit en alphabet arabe) possédant un contenu rédactionnel pertinent.

Afin de rendre les données recueillies sur les blogs exploitables, un certain nombre de traitements sont nécessaires. Nous les avons répartis comme suit :

- Transformation html vers texte ;
- Ajouter des marqueurs de phrases ;
- Sélection des phrases pertinentes ;
- Translittérer les mots étrangers écrits en script latin à des mots en script arabe ;
- Convertir les dates, les nombres, les pourcentages, les abréviations, les acronymes... ;
- Normalisation selon CODA ;
- Transcription des caractères spéciaux ;
- Suppression de la ponctuation.

En effet, lors du traitement de notre corpus, nous avons pu remarquer que certains traitements peuvent être considérés comme relativement indépendants de la langue, tandis que les autres traitements sont spécifiques et qui dépendant de la langue. Pour ce faire, nous construisons une boîte à outils générique qui contient des outils de traitement pour rendre notre corpus dialectal exploitable.

La figure 4.10 présente l'architecture de la boîte à outils générique développée dans notre travail. Après avoir déterminé tous les problèmes et traitements nécessaires pour l'obtention d'un corpus textuel, nous avons décidé de décomposer ces traitements en un ensemble de petits modules. Ensuite, ces petits modules sont répartis en deux groupes :

- les modules généraux, qui travaillent indépendamment de la langue ;

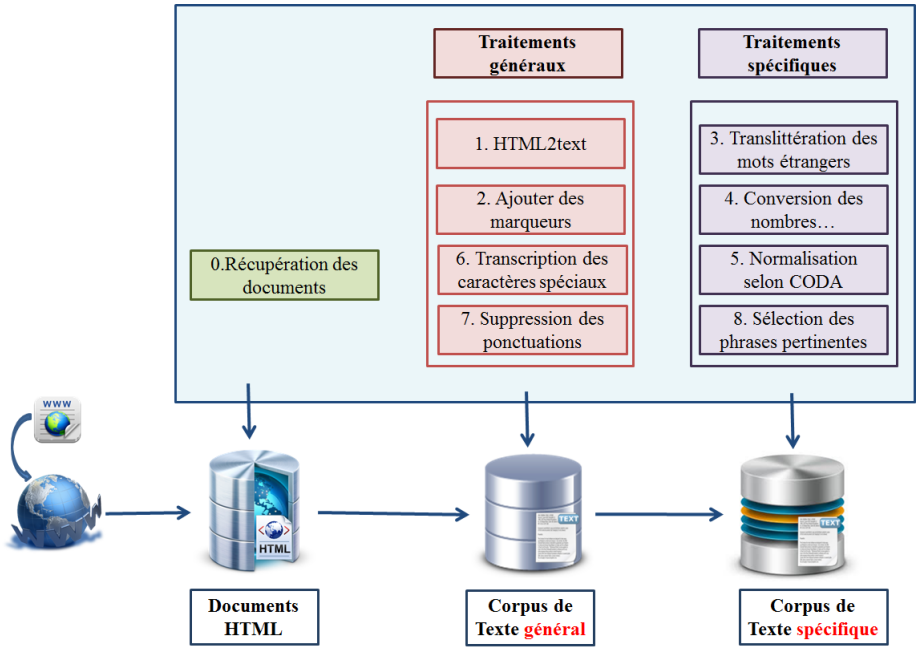

Figure 4.10 – L'architecture de la boîte à outils générique développée dans notre travail.

- les modules spécifiques, qui sont dépendants de la langue.

À l'issu de ces traitements, le corpus du texte construit comporte environ 18161 de mots.

4.7 Translittération des données en dialecte tunisien

Récemment, l'évolution des technologies de l'information et le développement des nouvelles formes avancées de communication ont influencé la communication entre les correspondants. Cette évolution a rendu la transmission de l'information plus facile et plus simple ce qui a engendré l'apparition de nouvelles formes d'écriture grâce à l'émergence de nouvelles formes avancées de communication tels que les emails, le chat, les SMS, les commentaires ... C'est pour cela nous avons recours à l'utilisation de ces potentiels écrits comme point de départ pour la construction de grands corpus d'une manière automatique. Toutefois, la majorité de ces messages et commentaires ont été écrits par les caractères de l'alphabet latin.

Ce fait est dû tout d'abord suite au manque des claviers arabes au début du développement des nouvelles technologies (PC, téléphones intelligents et tablettes) ce qui a conduit les Tunisiens à transcrire avec l'alphabet latin vu que ce dialectal

peut être écrit en utilisant les lettres arabes ou latines. En outre, c'est en raison de l'habitude et la facilité d'écrire en latin, surtout que les Tunisiens introduisent souvent dans leurs écrits et leurs conversations des mots en français (forme standard ou SMS). Cependant, le corpus écrit en caractère latin pose un problème pour le traitement des langues naturelles (NLP) du fait que certains outils sont devenus disponibles récemment pour traiter l'entrée du dialecte tunisien. Il nous faut donc un outil qui convertit le corpus écrit en caractère latin vers le script arabe. Tout au long de cette conversion nous avons suivi notre convention de normalisation CODA.

Nous présentons ensuite les différentes méthodes que nous avons utilisées pour construire notre corpus.

- La première méthode se base sur les SMS, nous avons demandé à la famille et les amis de nous envoyer leurs messages existants dans leurs téléphones mobiles. Le message le plus long est composé de dix mots, et le plus court se compose d'un seul mot.
- La deuxième méthode fait référence au Facebook qui est considéré comme le site de réseau social le plus populaire en 2013 selon le site « Le countries.com ». Comme ces réseaux sociaux jouent un rôle important dans la vie des Tunisiens, nous avons décidé d'utiliser leurs affichages, messages et commentaires écrits dans Facebook pour collecter notre corpus. Dans le but de maximiser la couverture du vocabulaire et de garantir la diversité de notre corpus, nous avons utilisé les différents types de pages Facebook (médias, la politique, les sports ...).
- La dernière méthode de collecte de données vise de s'occuper de Youtube qui est selon des études récentes comprend approximativement 20% de toutes les données HTTP, ainsi près de 10% de l'ensemble du trafic sur Internet. Dans le monde arabe, les gens ont recours à utiliser de plus les dialectes arabes (le dialecte égyptien, Golfe, tunisien, etc.) sur des sites comme Youtube pour écrire des commentaires et interagir avec leurs communautés. Toutefois, dans notre travail, nous avons conservé seulement les commentaires écrits par les utilisateurs en dialecte tunisien en caractères Latin.

4.7.1 L'orthographe spontanée du dialecte tunisien

À ce sujet, nous avons constaté qu'un mot en dialecte tunisien peut s'écrire en plusieurs façons puisque dans les cas où il n'existe pas d'orthographe standard, les gens utilisent une orthographe spontanée qui repose sur des critères différents et qui représente le critère principal celui de la phonologie. En effet, cette technique implique des mots écrits comme ils se prononcent autrement dit elle remplace un son avec une lettre latine ou un groupement de lettres latines. Cela dépend

principalement sur des hypothèses spécifiques à une langue à propos de la correspondance graphème-phonème. De surplus, une orthographe spontanée peut être influencée par des effets de la parole tels que l'étirement d'un mot (séquences répétées de lettres) pour exprimer des émotions intenses, par exemple, « Bnnnnina», «Mabrouuuk» et «Barrrrrrrcha » qui signifie بنينة [délicieux], مبروك [félicitations] et برشة correspond à [beaucoup].

Dans ce qui suit nous allons présenter certains aspects spécifiques de corpus dialectal écrit en caractère latin.

- **Les consonnes**

Nous présentons quelques équivalences entre l'alphabet latin et l'alphabet arabe extrait de notre corpus. Par exemple, les caractères latins « b », « s » et « l » sont utilisés pour représenter le son des lettres arabes ب /b/, س /s/ et ل /l/ respectivement. Néanmoins, nous avons rencontré quelques ambigüités en raison de l'absence de lettres latines assez suffisantes pour présenter toutes les prononciations des lettres du dialecte tunisien qui a été un obstacle dans le domaine de translittération. Par exemple, pour le caractère latin /t/ lui convient ت donc le caractère ط ne peut pas être présenté. De plus, il y a certains pairs de caractères latins qui peuvent être ambiguës dans leurs correspondances entre une lettre arabe simple ou paire de lettres : par exemple, « dh » peut être utilisé pour représenter les lettres ض /d/ et ده /dh/, et /kh/ peut être utilisé pour représenter خ et كه.

En outre, en cours de l'étude du corpus nous avons constaté que certaines lettres arabes étaient transcrites à travers des chiffres arabes. Ces chiffres peuvent remplacer des lettres et des sons qui n'ont pas d'équivalents dans l'alphabet latin. Par exemple, les chiffres 3, 5, 7 et 9 sont utilisés pour représenter le son des lettres ع /E/, خ /x/, ح /h/ et ق /q/ respectivement. Par ailleurs, lorsqu'un chiffre est suivi par « ‘ », les chiffres 3, 6, 7 et 9 changent leurs interprétations et deviennent غ /g/, ظ /d/, خ /x/, ض /d/. À cet égard nous notons que le recours aux chiffres est aussi une caractéristique du langage SMS français où les chiffres remplacent des séquences sonores reflétant la prononciation des chiffres, par exemple, « demain - 2m1 ». Cela entraîne des difficultés au niveau du déchiffrement des messages vu l'utilisation des chiffres dans les écritures en alphabet Latin.

- **Les voyelles**

Les voyelles sont utilisées dans l'écriture latine pour avoir les phonèmes arabes convenables avec une prononciation presque exacte. En effet, les Tunisiens utilisent les symboles en alphabet latin des voyelles qui sont comme suit (a, e, i, o, u, y) pour représenter les voyelles courtes et longues du dialecte tunisien.

- **Les mots étrangers**

Pour des raisons historiques, l'utilisation des mots étrangers est une caractéristique importante dans la communauté tunisienne qui est utilisée dans les conversation de tous les jours. De ce fait, il existe de nombreux mots étrangers utilisés et même intégrés dans les messages en dialecte tunisien comme « demain », «s alon », « hôtel », etc .

- **Les abréviations**

Cette technique consiste à réduire la taille des mots par l'élimination des voyelles tout en maintenant une liste des principales consonnes compréhensibles. Le corpus écrit en latin peut comporter certaines abréviations telles que «hmd», «wlh» et «slm» qui signifie الحمد الّله /Hamdallah/ [Merci Dieu], سلَام [la paix] respectivement.

- **Les effets sonores**

Nous avons également observé l'usage fréquent des observations écrites sur des effets de la parole, y compris les représentations du rire (par exemple, hhhhh), les pauses remplies (par exemple, euh), et d'autres sons (comme, hmmm).

- **Les acronymes**

Un acronyme est un sigle, formé des initiales d'un groupe de mots, formant une expression ou un nom d'une institution et se prononce comme un mot normal et non pas lettre par lettre. Par exemple, les acronymes «N» et «o/n » qui signifient respectivement نسمَة [nom d'une chaîne tunisienne] et نعم أو لَا [oui ou non].

4.7.2 Translittération vers le script arabe

Notre objectif est de sélectionner pour chaque mot écrit en caractère latin en entrée sa forme d'écriture arabe suivant CODA. Ceci est fait dans le but de générer automatiquement en premier un ensemble de translittérations possibles en écriture arabe en suivant CODA.

Dans le but d'accomplir la tâche de translittération, nous avons utilisé une approche basée sur les règles qui consiste à utiliser un ensemble de règles de translittération et un lexique d'exceptions. Ce lexique d'exceptions comporte essentiellement les abréviations et les acronymes. En effet, la forme de chaque mot exceptionnel est entrée avec sa forme écrite en caractères arabes. Le lexique d'exceptions est scanné en premier. Sinon, nous devons appliquer les règles au mot pour générer

sa forme arabe. Le processus de translittération se compose d'un certain nombre d'étapes bien définies :

D'abord, nous avons translitéré les abréviations et les acronymes à l'aide du lexique d'exception. Ensuite, dans la translittération toutes les émoticônes et les Emoji ont été remplacées par . Aussi, nous l'avons déjà indiqué, les gens répètent souvent des séquences de lettres pour exprimer des émotions intenses, de ce fait, nous avons supprimé toute répétition d'une lettre au-delà d'une répétition. Par exemple, nous avons transformé le mot « bninnnna » [délicieux] à « bnina». L'étape finale consiste à appliquer nos règles pour chaque mot. Étant donné que nous effectuons la translittération des mots écrits en caractère latin en écriture arabe suivant CODA, une phase de prétraitement est nécessaire. Par exemple, dans le cas où CODA exige un mot d'entrée divisé en deux ou plusieurs mots du script arabe, ainsi, nous indiquons ceci en ajoutant un tiret entre les mots. Par exemple, le mot écrit en latin « Ma5rajch » [Il n'est pas sorti] doit être divisé en deux mots arabes /Ma - 5rajch/ مَا ـ خرجش où مَا /équivalent de [pas] en français/ représente le clitique de négation en dialecte tunisien qui ne peut être attaché au mot suivant en fonction de CODA.

Comme mentionné ci-dessus, nous avons rencontré quelques ambiguïtés au niveau des consonnes dans l'écriture latine de dialecte tunisien à cause d'absence des lettres latines suffisantes pour présenter toutes les prononciations de l'écriture arabe, ce qui peut être un obstacle dans la translittération. De même, nous avons remarqué que seulement les experts du dialecte tunisien peuvent faire la distinction entre ces cas. Pour surmonter ces obstacles, nous avons proposé une solution qui consiste à énumérer toutes les versions possibles du mot en entrée. Après cela, l'utilisateur sélectionne le meilleur choix de toutes les possibilités. Par exemple, le mot écrit en latin « Hlal » contient le graphème latin « h » qui est utilisé pour représenter le son des lettres arabes ه /h/ et ح /H/. Ainsi, la sortie doit afficher toutes les possibilités de ce graphème هلَال /Hlal/ [croissant] ou حلَال /hlal/.

4.7.3 Evaluation de l'outil de translittération

Dans cette section, nous avons montré la qualité de translittération de notre corpus écrit en caractère latin vers le script arabe. À cause d'absence des outils automatique pour l'évaluation, nous avons demandé à des experts humains pour juger le degré de performance de notre script de translittération. Dans cette optique, nous avons effectué deux types d'évaluation : l'évaluation hors contexte et l'évaluation en contexte. Dans la section suivante, nous donnerons plus de détails sur les processus d'évaluation.

4.7.3.1 L'évaluation hors contexte

Nous avons demandé à des juges qui sont des locuteurs natifs du dialecte tunisien à translittérer manuellement un ensemble de 3.500 mots dans l'écriture arabe. Ces mots ne sont pas redondants. Cet ensemble de mots comprend en particulier des mots d'origine arabe et des mots d'origine étrangers tels que le français. À cet égard, l'évaluation consiste à comparer ce que notre système a proposé comme une translittération avec les décisions des juges. Partant de ce fait, nous calculons le rappel de notre système comme le pourcentage d'accord entre les transcriptions des juges et les transcriptions proposées par notre système. Le tableau 4.3 montre les résultats.

Type	***Rappel***
Les mots d'origine arabe	93%
Des mots étrangers	90%

Table 4.3 – Le rappel des translittérations des juges par notre système dans le cas d'évaluation hors contexte.

L'analyse a montré que les erreurs de mots d'origines arabes sont principalement dues à des raisons suivantes :

- Des erreurs dues à l'ambiguïté du mot écrit en caractère latin. Du fait que, l'entrée contient une faute de frappe, il rend impossible de produire la translittération correcte. Par exemple, l'entrée «5obs» contient une faute de frappe où le «s» final doit être «z» pour désigner le mot finale خبز /xubz/ [pain].
- Des erreurs se produisent lorsque le système génère la translittération de certains mots qui ne sont pas compatibles avec la forme CODA. Par exemple, le système génère la forme non-CODA لياَم /layyAm/ [les jours] au lieu de la forme correcte de CODA الءياَم /Alyyam/ [les jours].
- Autres types d'erreurs :
  - Des erreurs morphologiques : nous avons remarqué une translittération incorrecte de suffixe verbale de la troisième personne du pluriel وَا 'Wa' dans certains verbes. Par exemple, le système génère la forme verbale خرج /xarju/ [il sort] au lieu de la forme verbale correcte خرجوَا /xarjwA/ [ils sont sortis].
  - Des erreurs de segmentation : nous avons remarqué que certaines particules telle que لَا [ne] sont attachés aux mots. Par exemple, le système

génère la forme لَا مشَى /lAm$y/ [pas-de marche] au lieu de la forme correcte لَا مشَى /lAm$Y/ [non, il va].

- Des erreurs dues à la translittération incorrecte de certains mots étrangers. Par exemple, le système génère la translittération du mot étranger «courage» comme كورج /kwrAj/ mais selon les juges humains, ce mot doit être traduit comme كورَاج courage .

4.7.3.2 L'évaluation en contexte

À ce propos, nous avons demandé à 4 juges de translitérer 200 phrases contenant 832 mots. Notons que nous avons répété quelques mots dans le corpus de test, mais dans des contextes différents. Au début, nous avons testé les pourcentages d'accord entre les translittérations des juges. Le tableau ‘4.4 illustre les résultats de l'accord inter-juge. La variation de pourcentage est due au fait que, pour certains mots, les juges ne sont pas d'accord entre eux.

	2 juges	**3 juges**	**4 juges**
Accord	94%	93%	90%

Table 4.4 – Résultats de l'accord inter-juge.

Dans une analyse de l'accord inter-annotateurs, l'accord global entre les quatre juges était de 90%. Nous avons analysé tous ces désaccords et nous les classé en trois catégories de haut niveau :

- CODA : Certains cas de désaccord ont été liés à des décisions de CODA à cause de manque de connaissances des directives de cette convention
- Les mots étrangers : Certains cas de désaccord étaient liés à des mots étrangers. En fait, dans certains cas, les juges ne sont pas d'accord sur la translittération de mots étrangers. Par exemple, le mot français «demain» été translittéré en caractères arabes par deux juges comme دومَان /dwmAn/ [demain] et il a été translittéré en caractères arabes par deux autres juges comme دمَان /dmAn/.
- Ambiguïté : le désaccord des juges reflète une lecture différente de mot écrit en caractère latin qui a abouti à une caractéristique d'inflexion.

Après cela, nous avons effectué une deuxième évaluation qui a consisté à comparer ce que notre système a proposé comme translittération avec les propositions

des juges. Le pourcentage d'accord entre les translittérations des juges et les translittérations proposées par notre système a été calculé. Le calcul du pourcentage d'accord et de désaccord a été fait comme suit : s'il y a un accord entre la proposition de notre système et une seule proposition de l'un des quatre juges, nous avons attribué une valeur 1, dans le cas contraire, ca valeur devrait être 0. Le tableau 4.5 montre le pourcentage d'accord entre les translittérations des juges et les translittérations proposées par notre système dans le cas de l'évaluation en contexte.

Type	***Accord***
Les mots d'origine arabe	92%
Des mots étrangers	89%

Table 4.5 – Le pourcentage d'accord entre les translittérations des juges et les translittérations proposées par notre système dans le cas de l'évaluation en contexte.

Les erreurs sont principalement dues aux raisons suivantes :

- erreurs dues à l'ambiguïté de l'écriture du mot en caractère latin. Par exemple, le mot d'entré est montagne qui est dans le contexte "barcha jbAl" [beaucoup de montagnes], le système génère جبل /jbl/, tandis que la bonne réponse est جبَال [montagnes].
- Des erreurs se produisent lorsque le système génère des translittérations de mots qui ne sont pas compatibles avec la forme CODA.
- Des erreurs dues à la translittération incorrecte de certains mots étrangers.

4.8 Conclusion

Dans ce chapitre nous avons présenté un recueil des corpus pour de dialecte tunisien qui représente une tâche indispensable pour le développement d'un SRAP. Les données recueillies comportent d'une part des signaux de parole, et d'autre part des données textuelles. Ainsi, nous avons présenté les étapes de création de notre corpus nommé TARIC : Corpus de l'interaction des chemins de fer de dialecte tunisien dans le domaine de la SNCFT. En fait, la tâche essentielle de ce corpus d'audio consiste à demander des informations sur les services de chemin de fer dans une gare ferroviaire en dialecte tunisien. Le logiciel que nous avons utilisé pour la transcription est « Transcriber ». Tout au long de ce travail, nous avons adopté notre convention de normalisation CODA lors de la transcription de notre corpus.

Néanmoins, notre corpus TARIC que nous avons recueilli a atteint les 20 heures qui présentent une quantité limitée pour la conversion G2P. Dans le but de pallier le problème de la carence des données, nous avons essentiellement considéré des ressources issues à partir de deux méthodes de collection de données de grande quantité. Premièrement, une attention particulière a été apportée à une approche intéressante qui vise à «aspirer» les sites Web en dialecte tunisien. Deuxièmement, nous avons fait recours à l'utilisation d'un outil de translittération pour retranscrire les données recueillies en caractères arabe. Ainsi, il faut s'assurer que cette conversion est effectuée suivant la convention de l'orthographe CODA de l'arabe dialectal.

Dans le chapitre 5, nous présentons une approche pour la conversion G2P en vue d'obtenir un dictionnaire phonétique. Ce dernier est constitué comme un élément central de l'apprentissage des modèles acoustiques de SRAP.

Chapitre 5

Conversion G2P pour le dialecte tunisien

Sommaire

5.1 Introduction

Dans le cadre de la RAP, le système de conversion G2P permet de générer un dictionnaire de prononciation. Ce dernier est un élément central de l'apprentissage des modèles acoustiques. En fait, il s'agit d'associer chaque entrée du dictionnaire, qui est présentée sous la forme d'une séquence de graphèmes (i.e. chaque mot), à une suite de phonèmes qui lui est propre.

Dans ce chapitre, nous abordons la conversion G2P du dialecte tunisien en vue de la transcription automatique de la parole. Cette conversion ou phonétisation peut être définie comme la tâche de transformer un mot donné (séquences de graphèmes) à ses symboles phonétiques correspondants (séquences de phonèmes). Sa complexité varie selon la langue traitée. Par exemple, la conversion G2P de l'espagnol semble une tâche traitable en se basant sur des règles phonétiques simples et dépendantes de cette langue en raison de la correspondance plus ou moins directe entre l'écriture alphabétique et les systèmes phonétiques utilisés. Par ailleurs, il existe d'autres langues qui ont seulement des régularités partielles entre leur orthographe et les systèmes phonétiques comme le français ce qui engendre une ambiguïté dans la correspondance entre les systèmes orthographiques et phonétiques. Pour la langue arabe, la correspondance entre les systèmes orthographiques et phonétiques se situe entre le simple (espagnol) et le complexe (français).

Nous avons initié par un survol sur les problèmes de la conversion G2P du dialecte tunisien, nous visons dans ce chapitre de s'intéresser plus aux solutions proposées pour résoudre ces problèmes et les règles utilisées pour la tâche de conversion G2P de cette langue.

5.2 Les problèmes de conversion G2P du dialecte tunisien

La conversion G2P du dialecte tunisien est confrontée à de nombreux problèmes notamment les problèmes morpho-phonémiques, l'élision et également les variations phonologiques et phonétiques. De même, la Graphème en Phonème devrait tenir en compte l'apparition de nouveaux phénomènes comme l'assimilation et métathèses. Par ailleurs, cette conversion doit trouver des résolutions dans le cas où il existe des introductions de mots étrangers dans la langue et en cas d'irrégularités d'orthographe.

5.2.1 Le système d'écriture du dialecte tunisien

Le dialecte tunisien n'est pas considéré comme une langue indépendante comme le Français ou l'anglais, il présente une variété de MSA. De ce fait, nous ne pouvons

pas trouver une séparation stricte entre ce dernier et ce dialecte. D'ailleurs, il existe beaucoup de similitudes entre eux tels que le système d'écriture qui est de droite à gauche. Aussi, ce dialecte utilise le même alphabet que l'arabe avec quelques lettres supplémentaires qui sont obtenues à partir d'autres langues étrangères comme le français, le berbère et l'espagnol. De plus, les lettres de MSA et du dialecte tunisien sont reliées, même quand elles sont imprimées. D'une manière générale, chaque lettre peut apparaître en un maximum sous quatre formes différentes, selon son emplacement au début, au milieu ou à la fin du mot, ou même dans l'isolement.

Ce système d'écriture est composé en premier lieu par trois voyelles courtes, appelées signes diacritiques, figurant au-dessus ou en dessous d'une consonne. En deuxième lieu, il existe les trois symboles de Tanween qui sont Tanween Fatha, Tanween Kasra et Tanween Dhamma. Ces signes diacritiques se trouvent toujours à la fin du mot en représentant la combinaison d'une voyelle courte et le marqueur /n/. Néanmoins, ce type de signe diacritique est rarement utilisé en Dialecte Tunisien. Par contre, il est utilisé lorsque nous voulons montrer une politesse exagérée comme dans le mot أهلًا /bienvenue/ ou شكرًا /merci/. En fait, tous ces signes diacritiques mentionnés ont des divers objectifs. Ils modifient la valeur phonétique du graphème, aussi, ils permettent une lecture plus précise et évitent les ambiguïtés. En outre, leurs présences montrent si le texte est voyellé ou non, malgré que, de nos jours, la plupart des documents, textes et journaux ne sont pas voyellés.

En général, les 31 consonnes sont présentées en 28 graphèmes provenant de MSA et trois autres /P/, /V/ et /G/ se dérivent des mots étrangers. Aussi, ces consonnes sont réparties en des caractères orthographiques notamment Shamsi (Solaire) et Ghamari (Lunaire).

À part les voyelles courtes et les symboles de Tanween, nous pouvons constater l'existence des voyelles longues qui sont le prolongement phonétique des voyelles courtes et dont leurs formats sont les suivants :

- **Alif** : le prolongement phonétique de Fatha, son phonème est (AE :).
- **Waw** : le prolongement phonétique de Dhamma, son phonème est (UW).
- **Ya** : le prolongement phonétique de Karsa, son phonème est (IW).

Il faut noter que les voyelles longues «Ya» et «Waw» peuvent être utilisées comme des consonnes et prennent des voyelles courtes. Dans ce cas, elles sont appelées semi-consonnes, du fait qu'elles passent souvent d'une voyelle longue à une consonne. Contrairement au MSA, il existe en dialecte tunisien une nouvelle voyelle longue qui est un mélange entre Alef et Kasra afin de présenter la longueur phonétique de Kasra, son phonème est /EY/. Prenons à titre d'exemple le mot حرَام [Hre :m] [un vêtement traditionnel]. Encore, le système d'écriture de ce dialecte

comprend le symbole «Shadda» ou «gémination» qui apparaît généralement sur une consonne pour indiquer que le son de la consonne sera répété.

Il existe encore le symbole ligature qui est une association de deux symboles de caractères ou plus représentés par écrit par un seul symbole orthographique. Ces symboles de ligatures sont omni présents dans plusieurs langues, par exemple, la langue française connaît une ligature /œ/, appelé «e» dans l'«o» comme dans le mot (œuvre). Ainsi dans le dialecte tunisien, il y a six formes de ligature :

- Alif Lam (l'équivalent de «the» en anglais) comme le mot البيت /Albyt/ [the house],
- Lam et Alif لَا / lA/ comme le mot لَازم /lAzim/ [nécessaire],
- Hamza sur waw أُوء comme le mot مسءول [responsable],
- Hamza sur ya comme le mot مَاءدة [table],
- Hamza sur alif comme le mot أربعين [Quarante],
- Alif mada آ comme le mot أية [nom d'une jeune fille].

5.2.2 Les problèmes morpho-phonémiques

Tout comme dans la langue anglaise et la langue française, la conversion Graphème en Phonèmes du dialecte tunisien peut dépendre de la partie précédente et/ou la suite des mots. Dans ce qui suit, nous allons exposer les différents cas de dépendance. Premièrement, ce type de dépendance peut être vu dans les mots commençant par le préfixe Alef et Lam (l'équivalent de [the] en anglais) et suivi d'une lettre Shamsi. Dans ce cas, quand le mot débute par Alef et Lam et suivi par une lettre Shamsi et le mot qui précède est terminé par une «Sukun», comme dans cette expression حلت الشمس /hliT Al$ms/ [le soleil brillait], le préfixe Alef et Lam se prononce comme إِ /Ae/, la lettre Shamsi suivante n'est pas géminée et le préfixe est mélangé avec le mot précédent. Si le mot débute par Alef et Lam, suivi par une lettre Shamsi et le mot précédent se termine par une voyelle longue, comme cette expression حلوَا الشبَاك /hlwA Al$bAk/ [Ils ont ouvert les fenêtres], le préfixe Alef et Lam est omis et la lettre Shamsi n'est pas géminée. Ainsi, ces deux mots ne sont pas mélangés. Deuxièmement, nous pouvons voir cette dépendance lorsque nous utilisons le clitique de négation du dialecte tunisien qui a ce format : مَا + تمشي + ش /mA+ tm$y + $/ [ne va pas]. Dans ce cas, la négation مَا /mA/ [pas] est attachée au mot suivant et devient juste une courte consonne م /ma/.

Troisièmement, cette dépendance est présente lorsque les pronoms démonstratifs هذَا /ha*A/ [ce] et هذي /ha*y/ [cette] sont utilisés avant un nom contenant Alef et Lam. Dans ce cas, ces pronoms sont réduits à une consonne et rattachés au nom. Par exemple, dans l'expression هذَا الطفل /h*A AlTfl/ [ce garçon], le pronom هذَا /ha*A/ [ce] sera prononcé comme une consonne ه /h/ et l'expression sera هَالطفل /hAltfl/. Quatrièmement, la préposition من /min/ [de], lorsqu'elle est utilisée avec un nom commençant par Alif et Lam, elle sera réduite à une consonne et attachée au nom, comme dans l'expression من الدَار /min dAr/ [de la maison] qui deviendra مَالدَار /mi dAr/. Cinquièmement, au niveau de la coordination conjonction مع /mE/ [avec], lorsqu'elle est utilisée avec un nom commençant par Alif et Lam, elle sera réduite à une consonne et attachée au nom. Notamment cet exemple, مع بعضنَا /mE bEDnA/ [ensemble] qui deviendra مبعضنَا /mbEDnA/. Sixièmement, dans le cas de la préposition في /fy/ [dans], lorsque cette dernière est utilisée avec un nom commençant par Alef et Lam, la dernière longue voyelle devient une voyelle courte. Cette préposition sera mélangée avec le mot suivant. Par exemple, في الصبَاح /fy sbAh/ [dans la matinée] deviendra فَالصبَاح /fisbAh/. Septièmement, la préposition إلَى /AlA/ [à], lorsqu'elle est utilisée avec un nom commençant par Alef et Lam, elle sera réduite à une consonne et mélangée avec le mot suivant. Par exemple, ألَى البيت /AlA Albayt/ [à la maison] deviendra لبيت /lilbyt/. Huitièmement, la préposition كيف /kyf/ [comme], lorsqu'elle est utilisée avec un nom commençant par Alef et Lam, elle sera réduite à une consonne et la voyelle longue devient une voyelle courte. Cette préposition sera mélangée avec le mot suivant. Par exemple, كيف العَادة /kyflEAda/ [comme d'habitude] deviendra كَالعَادة /KAlEAdt/ . Finalement, la préposition علَى /ElA/ [sur], lorsqu'elle est utilisée avec un nom commençant par Alef et Lam, elle sera réduite à une consonne ع /E/ et la dernière longue voyelle ى /y/ devient une voyelle courte /a/. Cette préposition sera mélangée avec le mot suivant. Par exemple, علَى الطَاولة [sur la table] deviendra عَالطَاولة /EyAwlt/ .

Il existe d'autres problèmes rencontrés en dialecte tunisien que nous allons présenter. Parmi ces problèmes, nous avons les problèmes d'élision.

5.2.3 Les problèmes d'élision

L'élision désigne la suppression d'une voyelle à la fin d'un mot avant la voyelle de départ du mot suivant. C'est un problème rencontré dans la langue française avec la prononciation du graphème «e» qui est parfois supprimé et devient un phonème vide. Néanmoins, l'élision n'est pas limitée seulement à la langue française, mais elle peut être rencontrée en arabe et en italien. Effectivement, dans le dialecte tunisien, les problèmes d'élision ont été trouvés lorsque il y a le graphème Alef, le graphème Hamza et le graphème celui de Ta-marbouta.

Par ailleurs, le graphème Alef peut être soit silencieux ou réalisé comme le son d'une voyelle longue en fonction de son emplacement dans le mot. D'après ce qui précède il s'agit de trois cas de prononciation pour le graphème Alef.

- En premier lieu, il est silencieux lorsqu'il se trouve à la fin du mot et apparaît dans le morphème وَا /wwA/ qui indique une conjugaison masculine plurielle dans les verbes.
- En deuxième lieu, lorsque ce graphème Alef est une partie du Tanween-Fatha, il est également silencieux. Par exemple, dans le mot شكرًا /$krAn/ [merci], le graphème Alef est supprimé.
- En dernier lieu, quand il est au milieu du mot comme دَار /dAr/ [maison], le graphème Alef est employé pour prolonger phonétiquement le graphème د /d/.

Au surplus, nous avons le graphème Hamza ا /A/ qui peut être vu en deux états :

- D'une part, il est silencieux quand il vient à la fin du mot, par exemple, le mot هوَاء /hwA/ [l'air].
- D'autre part, au début du mot, le graphème Hamza est prononcé, à l'exception lorsque le mot contenant ce graphème Hamza est se situe au milieu d'une phrase, il n'est pas prononcé.

En outre, le graphème Ta-marbouta qui est toujours situé à la fin de certains mots représentant la marque du féminin.

- Il peut être soit silencieux ou réalisé dans le son pour /t/ : Par contre, il n'est pas silencieux quand le mot suivant commence par Alef et Lam et nous donne le son de /t/. Par exemple, ليلة السبت /liltsibt/ [samedi soir].
- Encore, il n'est silencieux dans l'autre cas lorsque le mot suivant ne commence pas par Alef et Lam.

5.2.4 Les variations phonétiques et phonologiques

Comme nous l'avons déjà signalé, parmi les spécificités remarquables du dialecte tunisien, nous notons la présence des mots empruntés de différentes langues tels que le français, le berbère, l'italien, le turc, l'anglais et l'espagnol. La présence de ces mots est le fruit de nombreux facteurs et des événements historiques survenus tout au long des siècles, tels que : les invasions islamiques, la colonisation française, les migrations, les échanges commerciaux, etc. Or, beaucoup de mots empruntés sont utilisés dans le discours du peuple tunisien sans être adapté à la phonologie tunisienne, ce qui provoque l'introduction de certaines lettres étrangères comme /P/, /V/ et /G/.

Aussi nous avons constaté plusieurs variations phonologiques particulières dans le dialecte tunisien. Comme le MSA, un graphème peut correspondre à des graphèmes différents : les trois graphèmes ا ،و ،ي correspondent soit à des voyelles longues respectivement /A/ /w/ /y/, s'ils ne portent pas des voyelles courtes, soit des consonnes, respectivement /a/ /w/ /y/ s'ils portent des voyelles courtes. À titre d'exemple, le graphème est transcrit phonétiquement /w/ dans le mot ورد /ward/ (fleurs] et /w/ dans le mot تونس /twns/ [Tunisie]. Plusieurs graphèmes peuvent correspondre à un seul phonème. Comme le cas des deux graphèmes et qui correspond au phonème /t/. Il existe une variation de la prononciation de certaines consonnes et voyelles, et parfois de nouveaux phénomènes apparaissent, notamment, l'utilisation de la consonne occlusive sourde de MSA ق [q] qui a une double prononciation. Dans les dialectes rurales, elle est prononcée ڤ /G/ et dans les dialectes urbains, elle est prononcée ق /q/. C'est une propriété qui reflète une coupure socio-géographique entre les parlers urbains et les parlers ruraux. Cependant, certains mots comme le mot بقْرَة /baQra/ [vache] est toujours prononcé /bagra/ quel que soit la région. La consonne de MSA ض /*/ qui peut avoir plusieurs prononciations possibles telles que ض /*/ ou ذ /D/ ou د /d/. Par exemple, le mot مَاضي /mA*y/ dans l'expression سَاعة مَاضي [13 heures] est prononcé مَاضي /mA*y/ ou مَاذي /mADy/ ou مَادي /mAdy/. La consonne س /s/ peut être prononcée comme س /s/ ou ص /S/ . Par exemple, le mot رسول /raswl/ [prophète] est prononcé رسول /raswl/ ou رَسْول . La consonne ش /$/ peut être prononcée comme س /s/ dans certains cas. Par exemple, le mot شجرة /$jrat/ [arbre] est prononcé شجرة /$jrat/ ou سجرة /sjrat/ . La consonne ض peut être prononcée comme ض /D/ ou ظ /DD/. Par exemple, le mot نفترض /niftariD/ [supposons] est prononcé نفترض

/nftariD/ ou نفترظ /nftariDD/. Dans certains mots tels que ثمة /vamma/ [exister], la consonne de MSA ث /v/ peut se prononcer en deux façons : ث /v/ ou ف /f/. La consonne de MSA ط /T/ est parfois prononcée ط /T/ et d'autres moments ت /t/. Par exemple, أعطيني /AETyny/ [donnez-moi] est prononcé أعطيني ou أعتيني . La consonne ع /E/ est parfois prononcée ع /E/ et parfois ح /H/. Par exemple, متَاعهَا /mtAeHA :/ [la sienne] est prononcé متَاعهَا ou متَاحهَا . La consonne أ /A/ , lorsqu'elle est située au milieu d'un mot, elle est parfois prononcée أ comme dans سءوَالَات [questions], parfois prononcée ي /y/ comme dans وسَاءل /wasa :yil/ et dans d'autres cas, elle est prononcée ه /h/ ou أ comme dans le mot نسءلك /ni-shalik/. Dans le mot de le MSA فزدق /fuzdaq/ [Pistache] est prononcé en dialecte tunisien فستق /fustaq/. Nous remarquons que la seconde consonne ز /z/ devient س /s/ et la troisième consonne د /d/ devient ت /t/. Au début du mot, le Hamza est parfois prononcée de différentes manières. Si le mot est au début de la phrase, le coup de glotte est prononcé. Si le mot est dans le milieu de la déclaration, le coup de glotte est omis. Un nouveau phonème /EY/ est apparu en dialecte tunisien, comme dans le mot حرَام [tabou].

Nous relevons parfois l'ajout du phonème إِ /Ai/ pour soutenir la prononciation de la première consonne silencieuse d'un mot. Autrement dit, l'alif prosthétique est rétabli devant les noms dont la première consonne est menacée de devenir quiescente (que l'on ne prononce pas) : Par exemple, le mot مسَاكن /Mseken/ [nom de ville Tunisienne] peut être prononcé comme أمسَاكن /Ei Mseken/. L'alif prosthétique avec attaque vocalique qui est généralement supprimé là où il devait normalement exister. Par exemple, أبو بكر /Abw Bakr/ [nom d'un homme] est prononcé comme بو بكر /bw Bakr/, dans ce cas l'alif au début de mot devient silencieux. Nous avons remarqué que l'apparition du phénomène métathèses qui est défini comme l'échange entre deux phonèmes en contact ou à proximité. Par exemple, شمس /$ams/ [soleil] est prononcé سمش /sam$/ ou شمس /$ams/. En effet, un autre phénomène est apparu qui est l'assimilation phonétique. Ce phénomène se définit comme suit : action lorsque un phonème (l'élément de l'assimilateur) communique une ou plusieurs de ses caractéristiques à un phonème voisin (l'élément assimilé).En dialecte tunisien, le phonème ج se transforme au phonème ز /z/. Par exemple, le mot جرجيس /jrjys/ [nom d'une ville tunisienne] devient جرجيس

/jrjys/ ou زرزيس /zrzys/. Dans un autre exemple, le mot عجوز /Ejwz/ [un vieil homme] devient عجوز /Ejwz/ ou عزوز /Ezwz/. Nous avons observé l'élimination d'un ou deux consonnes dans quelques mots.Par exemple, قلت لك /qiltlik/ [je vous ai dit] peut être prononcée قتلك /qitlik/, nous avons constaté que la consonne ل /l/ est éliminée. Dans d'autres exemples comme le mot مَا نعرفش /mA naEraf$/ [je ne sais pas] peut être prononcé مَا نعش /ma : naEr$/, nous avons remarqué que les deux consonnes ر /r/ et ف /f/ sont éliminées. Nous avons remarqué la tendance à remplacer des voyelles courtes par des voyelles longues. Par exemple dans le mot فرنسي [Français] est prononcé فرَانسي. Nous avons remarqué la transformation de certaines voyelles longues en des voyelles longues. À titre d'exemple, le mot أيطَالي /ATAly/ [Italien] est prononcé أطلي. Il faut indiquer que les numéros en dialecte tunisien ont des caractéristiques spécifiques : Les numéros entre «trois» et «neuf» acceptent un double prononciation, dont l'un subit l'élimination de certaines consonnes et une modification de la voyellation. Par exemple, le nombre تسعة /tsEat/ «neuf» peut être prononcé sans aucuns changements quand il est isolé (il n'existe pas de mot après). En revanche, quand il ya un mot après ce nombre comme dans l'expression تسعة ميّة [l'argent], il est prononcé comme تسع /tsaE/. Ainsi, les numéros à partir de «onze» acceptent également deux prononciations et nécessitent l'ajout le phonème «n» à la fin du numéro. Par exemple, le nombre ثنَاش /vnA$/ [douze] peut être prononcé sans aucuns changements quand il est isolé (il n'existe pas de mot après), par contre, quand il ya un mot après ce nombre comme dans l'expression ثنَاش ألف /vnA$/ [l'argent], il est prononcé comme ثنَاش + ن /vnA$ n/.

Dans les sections suivantes, nous allons présenter le processus de conversion G2P du dialecte tunisien et les différentes règles phonétiques que nous avons dégagées à partir de notre corpus. Ensuite, nous allons donner des solutions pour résoudre les problèmes de cette conversion présentée dans les sections précédentes. De plus, nous allons montrer la manière d'application de ces règles pour effectuer la tâche de phonétisation dans les deux cas de : mots voyellés ou non voyellés.

5.3 La conversion G2P : approche à base de règles

Rappelons que la conversion G2P se base, comme nous comme nous l'avons déjà présenté précédemment dans le chapitre 3, sur une approche manuelle, approche

à base de règles ou approche guidée par les données. Dans ce travail nous avons opté l'approche à base de règles. Pour ce faire, nous avons identifié un ensemble de règles de prononciations. En fait, l'élaboration et la conception de ces règles ont été modélisées par des locuteurs natifs et des experts linguistes de cette langue. Nous avons remarqué qu'il y a des mots qui ne peuvent pas suivre notre ensemble de règles phonétiques. Ainsi, il est nécessaire de définir un lexique d'exceptions.

Avant de présenter les règles de conversion G2P du dialecte tunisien nous allons traiter quelques exemples du lexique d'exceptions.

5.3.1 Le lexique des exceptions

L'orthographe du dialecte tunisien est assez régulière, hormis dans certaines situations où certains mots violent les règles de prononciation régulières. Autrement dit, la prononciation de ces mots ne correspond pas à leurs graphies. Pour pallier ces formes irrégulières, le recours à un dictionnaire de lexique d'exceptions s'avère nécessaire. En effet, la forme phonétique de chaque mot exceptionnel est entrée avec sa forme graphémique. Ainsi, ce lexique d'exceptions est scanné en premier lieu. Dans le cas contraire, il faut appliquer les règles au mot pour générer sa forme phonétique. Le lexique d'exceptions est subdivisé en plusieurs catégories prédominantes :

La première catégorie inclut les pronoms démonstratifs du dialecte tunisien tels que هذَا /ha*A/ [ce mâle], هذي /ha*y/ [cette femelle] ou هذوم [ces femelles]. Les exceptions de ces pronoms résident d'une part, dans la génération de la voyelle longue /AE :/ pendant la prononciation, qui ne figure pas dans la forme orthographique du mot, autrement dit, l'allongement de la première consonne « h ». D'autre part, le raccourcissement du mot en une seule lettre « h », comme illustre l'exemple du pronom هذَا /ha*A/ [ce] qui accepte deux prononciations, la première met l'accent sur l'allongement de la consonne « h » d'où la prononciation de ce pronom devient هَاذَا /hA*A/. La deuxième prononciation se réduit à une seule lettre « h » comme dans le mot هَالطفل /hA/ [ce garçon]. La deuxième catégorie comporte certains noms exceptionnels comme les noms d'Allah الّله [dieu]. Cette catégorie a le même problème que le pronom personnel (la génération de la voyelle longue /AE :/ pendant la prononciation). La troisième catégorie comprend les pronoms personnels du dialecte tunisien comme أنَا /AnA/ ou أني qui signifient [moi] dans les régions du sud et du nord de la Tunisie respectivement. Les exceptions de ces pronoms personnels se trouvent dans le remplacement de la Hamza avec le Hamzet wasl au cours de la prononciation. De même, il existe une autre exception qui peut apparaitre dans la prononciation du pronom personnel أحنَا /AhnA/

[nous]. Généralement, ce pronom est prononcé dans la région du nord أحنَا et dans la région du sud نحنَا /nhnA/, dans ce cas il s'agit du remplacement de la première consonne أ par la consonne ن. La quatrième catégorie, est la plus large, elle englobe plusieurs types d'exceptions de mots par rapport aux autres catégories. Elle se caractérise par la présence des exceptions variées qui ne partagent pas le même type. Par conséquence, à l'aide de cette catégorie, nous avons pu résoudre les problèmes morpho-phonémiques et les problèmes de variations phonétiques et phonologiques du dialecte tunisien décrits respectivement dans la section 6.1.3 et 6.1.5. En fait, cette catégorie est subdivisée en plusieurs sous catégories. Traitons la première sous-catégorie contenant généralement les mots qui subissent quelques modifications telles que l'élimination de certaines consonnes quand elles sont prononcées. Par exemple, le mot قتلك /qiltlik/ [je vous ai dit] peut être prononcé قتلك /qitlik/, nous constatons que la consonne ل /l/ est éliminée. Dans ce cas, nous avons mis une double prononciation de ce mot. Ensuite, la deuxième sous-catégorie intègre les mots où le phénomène du Métathèse apparaît. Sur ce point, nous avons collecté depuis notre corpus ce type d'exception ainsi, notre base contient cette exemple de mot suivant le mot شمس /$ams / [soleil]. Encore la troisième sous-catégorie s'occupe d'un autre phénomène qui est l'assimilation. Nous pouvons tirer quelques exemple de mots de ce phénomène à partir de notre base d'exception comme le mot جرجيس /jrjys/ [nom d'une ville tunisienne] devient جرجيس ou زرزيس /zirzys/. Finalement la quatrième sous-catégorie s'intéresse aux mots qui subissent un changement des voyelles longues en des voyelles courtes. Comme le mot : la préposition في /fy/ [dans], dans ce cas cette préposition a deux prononciations possibles (في ou ف) etc.

En règle générale, le processus de phonétisation des mots s'effectue tout d'abord en consultant la base du lexique d'exceptions puis en appliquant les règles de phonétisation. De ce fait, dans la section qui suit nous allons présenter ces règles en précisant leurs caractéristiques spécifiques avec des exemples illustratifs.

5.3.2 Les règles phonétiques du dialecte tunisien

5.3.2.1 Format des règles

L'objectif principal de nos règles consiste à convertir une séquence de caractères orthographiques (graphèmes) à une séquence de phonèmes. En dialecte tunisien, les règles sont fournies pour chaque lettre. Chaque règle essaie de faire correspondre certaines conditions relatives au contexte de la lettre et de fournir un remplaçant. Les règles se lisent de droite à gauche comme indiqué la figure5.1 :

Figure 5.1 – Le Format des règles de prononciation.

Cette règle désigne que :

- **Graphème** : est la lettre courant dans le mot.
- **Condition droite** : possède l'un des formats suivants :
  - $<? \Leftarrow Patron >$: le contexte avant la position courant «Graphème» doit être considéré.
  - $<? <!Patron >$: le contexte avant la position courant «Graphème» ne doit pas être considéré.
- **Condition gauche** : peut prendre l'une de ces deux formats :
  - $<? = Patron >$: le contexte après la position courant «Graphème» doit être considéré.
  - $<!Patron >$: le contexte après la position courant «Graphème» ne doit pas être considéré.
- **Phonétisation** : est soit un phonème ou séquence de phonèmes ou nulle (*) si le graphème est omis dans la prononciation.

5.3.2.2 L'application des règles

Le processus de phonétisation pour les mots voyellés et ceux non voyellés se déroule en deux phases à savoir, la consultation de la base du lexique d'exceptions et l'application des règles de phonétisation.

Rappelons qu'il existe plusieurs mots qui ne se lisent pas selon des règles d'écriture bien déterminées. De ce fait, il a fallu les intégrer dans un lexique d'exceptions. En effet, ce lexique est consulté avant l'application des règles. Si le mot est parmi les exceptions, il est encodé directement sous forme phonétique. Sinon, nous appliquons les règles pour générer sa forme phonétique. Pour convertir la séquence de graphèmes en séquences de phonèmes, nous avons commencé par découper chaque mot en lettres. Ainsi, l'application de règles est décrite comme suit : La première règle trouvée qui satisfait les trois conditions sur Graphème, Condition droite et Condition gauche est immédiatement appliquée. Dès qu'une règle a été appliquée, on réitère le processus pour phonétiser la suite du mot. Nous obtenons ainsi la chaine phonétique (la prononciation) de chaque mot en concaténant les phonèmes

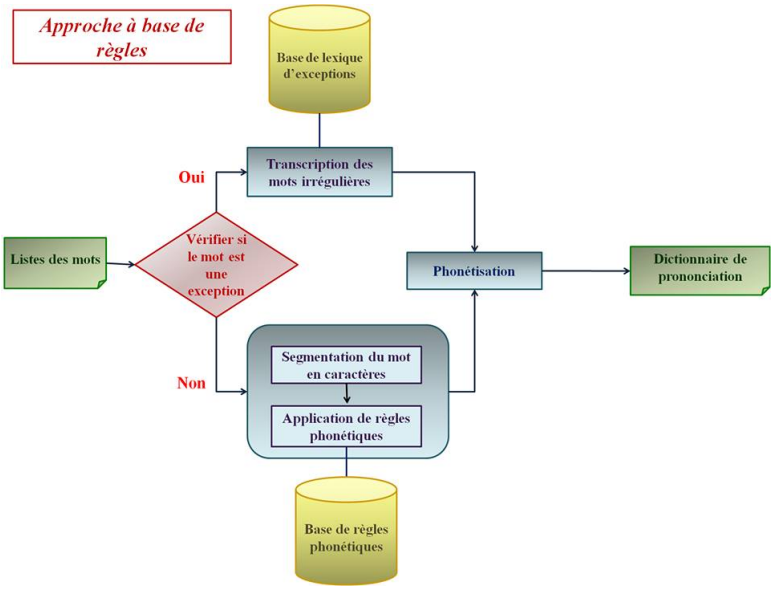

FIGURE 5.2 – Les étpes de la conversion G2P à base de règles

obtenus. La figure 5.1suivante illustre le processus de conversion graphème en phonème que nous avons suivi :

Comme nous avons mentionné dans les sections précédentes, le dialecte tunisien a trois voyelles courtes, autrement dit des marques diacritiques figurant ci-dessus ou en-dessous de la consonne. De ce fait, l'objectif principal de ces signes diacritiques est de faciliter la lecture et à rendre le texte moins ambigu. En fait, l'absence des voyelles courtes engendre l'ambiguïté au niveau phonétique et par conséquent au niveau sémantique. Par exemple le mot كتب /ktb/, peut avoir différentes transcriptions phonétiques comme كِتِب /ktib/ [il a écrit], كُتُب /ktub / [livre]. Ainsi, la détermination des formes phonétiques des mots voyellés est différente de celle des mots non voyellés et l'ordre d'application des règles phonétiques n'est pas similaire dans les deux cas. Nous allons présenter par la suite la différence dans la détermination des formes phonétiques pour les mots voyellés et les mots non voyellés et les modalités utilisées pour l'application des règles phonétiques dans les deux cas.

- **Application des règles pour les mots voyellés**

Le premier traitement pendant la phase d'application des règles consiste à segmenter un mot en une séquence de caractères. Ainsi, chaque mot est segmenté en consonnes et voyelles. Une fois cette segmentation a été effectuée, nous devons préciser le type pour chaque caractère du mot en vue de faciliter l'application des

règles. Tel que mentionné avant, il existe deux types de consonnes : les consonnes Shamsi et les consonnes Ghamari. En outre, il existe deux types de voyelles : les voyelles courtes et les voyelles longues. Aussi, il y a trois lettres ا, ي et و, qui peuvent être soit des voyelles longues ou consonnes.

Après la segmentation et l'identification de type de chaque caractère du mot, l'application des règles phonétiques s'effectue dans le sens de la lecture du mot (de droite à gauche), autrement dit, il commence par la première lettre du mot ainsi l'ordre des lettres est respecté. L'application de ces règles est faite conformément à un ordre bien déterminé, les règles de graphèmes et les règles de lettres étrangères sont appliquées en premier lieu. En second lieu, les règles d'élision et Sukun sont appliquées. En fait, les règles présentes produisent des sons nuls et peuvent être considérées comme le prétraitement pour la génération de la conversion Graphème en Phonème avant l'application de l'ensemble ordonné des règles. Ensuite, les règles d'élision et Sukun sont suivies par les autres règles non ordonnées telles que la règle Ya-Maqsoura, les règles Tanween, la règle de la gémination et les règles de ligature. Les règles phonétiques restantes sont commandées et appliquées dans l'ordre suivant :

- Les règles Shamsi et Ghamari sont appliquées par la suite.
- Les voyelles longues sont générées suivant le contexte des voyelles courtes.
- Les graphèmes des voyelles courtes restants sont remplacés par leurs prononciations respectives.

Après avoir traité la détermination des formes phonétiques pour les mots voyellés nous passons par la suite à explorer la détermination des formes phonétiques pour les mots non voyellés.

- **Application des règles aux mots non voyellés**

Généralement, la présence de corpus textuel en dialecte tunisien avec des mots qui portent des voyellés est presque absente et très limitée. Ainsi, vu l'absence des voyelles courtes en dialecte tunisien, une ambigüité dans la lecture et même la sémantique du mot est engendrée. Cette ambigüité apparait dans la détermination de la forme phonétique d'un mot qui est devenue plus difficile par rapport au mot voyellé. Afin de traiter cette ambiguïté dans la langue arabe, beaucoup de travaux de recherche ont été proposés dans la communauté du TALN dans le but de pallier le manque des voyelles tels que les travaux de voyellation des textes MSA [Roth et al., 2008 ; Elshafei et al., 2006]. Néanmoins, les outils fournis pour le MSA ne peuvent pas traiter le dialecte vu la différence morphologique, syntaxique, lexicale et phonétique entre la langue arabe et les dialectes. De ce fait, en dialecte tunisien le problème de l'absence des outils de voyellation automatique n'est pas encore

résolu. Ainsi, il faut trouver une solution pour avoir la prononciation exacte des mots non voyellé. Par conséquence notre problématique primordiale consiste à déterminer l'emplacement des voyelles courtes.

- **Principe d'emplacement des voyelles courtes dans les mots non voyellés**

À cet égard, lorsque nous avons collecté notre corpus TARIC, nous avons essayé de mettre les voyelles courtes lors de la transcription manuelle. Ensuite, d'après l'étude phonétique que nous avons effectuée de ce corpus, nous avons dégagé ces trois principes : Le premier principe montre qu'un mot en dialecte tunisien se termine soit par une consonne silencieuse (avec Sukun) ou une voyelle longue. Le deuxième principe prouve que chaque voyelle longue ou voyelle courte est toujours précédée et suivie par une consonne silencieuse (sans voyelles courtes). Autrement dit, nous ne pouvons pas trouver deux voyelles longues ou courtes successives. Finalement, le troisième principe indique qu'un mot ne peut pas avoir deux consonnes successives qui portent une voyelle longue ou courte sauf les mots avec «Shadda». C'est à dire, lorsqu'il s'agit d'un mot avec shadda, la consonne qui porte shadda et la consonne qui lui précède doit porter aussi soit une voyelle courte soit une voyelle longue.

Par voie de conséquence, ces trois principes ont servi à déterminer l'emplacement des voyelles courtes.

- **Autres règles de phonétisation pour les mots non voyellés**

Outre ces trois principes nous avons remarqué qu'il y a d'autres règles peuvent aussi nous faciliter de déterminer l'emplacement des voyelles courtes. Nous allons présenter dans ce qui suit ces différents cas :

- Si les deux semi-consonnes و et ي sont placées au début d'un mot, elles sont toujours des consonnes Par exemple : le mot يعطيك /Y AE AI TT IY K/, ي est une consonne. le mot وقتَاش /W AE Q T AE : SH/, و est une consonne.
- Si elles sont situées au sein d'un mot, elles peuvent être soit des voyelles longues soit des consonnes. Dans ce cas nous devons donner une double phonétisation de ces semi-consonnes. Par exemple dans le mot شلوقت peut avoir deux prononciations : /SH L AE W Q T/ ou /SH L UW Q T/, où و peut être une consonne ou une voyelle longue.
- Si elles se trouvent à la fin d'un mot, elles peuvent être encore soit des voyelles longues soit des consonnes. Par exemple dans le mot سَامحني /S AE : M AE

HH N IY/, ي est une voyelle longue. Dans un autre exemple de mot جي, le ي est une consonne.

En traitant la lettre و nous avons remarqué qu'il existe des cas particuliers comme dans cet exemple : وَاو الجمَاعة : و + ا dans l'exemple de mot خرجوَا, ainsi و est suivie par Alif à la fin du mot d'où elle est une voyelle longue. Un autre exemple : dans le mot مشَاو, le ww> n'est pas suivie par aucune lettre à la fin du mot et précédé par « Alif » alors elle est une consonne. De même, dans le cas où la lettre و est suivie par Ta-marbouta ة, elle est par la suite une consonne qui porte soit la voyelle courte Fatha soit la voyelle courte Kasra comme dans cet exemple : كسوة /K IH S W AE/ ou bien غدوة /GH UH D W AE/. De plus, lorsque و porte Shadda et suivie par ي donc le و est une consonne et ي est une voyelle longue comme dans le mot متلوّي [nom de ville].

- Au surplus, il existe des cas spécifiques pour la lettre ي lorsqu'elle se trouve à la fin d'un mot et précédée par ا elle devient une consonne comme dans ce cas مَاي /M AE : Y/. Encore, quand ي porte Shadda et suivie par ا, le ي est une consonne comme dans l'exemple ميَّات /M IH Y Y AE : T/. Par ailleurs, si dans un mot on trouve la lettre ي suivie par une autre lettre ي alors la première lettre est une consonne et la deuxième est une voyelle longue comme dans ce cas رَايّ /R AE : Y IY/.
- En fait, nous pouvons dégager d'autres règles telles que celles du Ta-marbouta ة qui ne peut pas être silencieuse lorsqu'elle est précédée par la voyelle courte Fatha ou bien silencieuse si elle est précédée par la voyelle courte Kasra. Par exemple :
  - Le mot مدرسة : Ta-marbouta est précédé par la voyelle courte Fatha d'où elle n'est pas silencieuse.
  - Le mot جمَاعة : Ta-marbouta est précédé par la voyelle courte Kasra d'où elle est silencieuse.
- En outre, s'il existe au début d'un mot لِل il faut insérer la voyelle courte Kasra entre les deux lettres ل notamment cet exemple لِلعيد /L IH L AI IY D/.

D'après ces trois principes et les règles de prononciation que nous avons présenté avant, nous avons pu distinguer quatre types de voyelleation de mots non voyellés en dialecte tunisien :

1. Le premier type collecte les mots contenant des voyelles longues (Alif ا Waw و - Ya ي).
2. Le deuxième type est celui des mots qui portent Shadda où il s'agit du doublement de consonne.
3. Le troisième type regroupe les mots qui possèdent des voyelles longues et Shadda en même temps.
4. Et finalement, les mots simples qui ne contiennent ni Shadda ni voyelles longues, leurs voyelles courtes sont placées selon le nombre des consonnes. De ce fait, nous devons prendre en considération le nombre des consonnes pour savoir l'emplacement des voyelles courtes. Par exemple si le mot simple comporte deux consonnes la voyelle courte doit être placée au milieu.

- **Ordre d'application des règles**

Dans ce cas, le premier traitement dans l'application des règles consiste à segmenter le mot en caractères et cette segmentation touche seulement les consonnes vu l'absence de voyelles courtes et de signes de Tanween. Contrairement au cas des mots voyellés, l'application des règles phonétiques des mots non voyellés ne respecte pas un ordre bien défini. De ce fait, la conversion G2P ne suit pas le sens de lecture du mot.

En effet, nos règles phonétiques pour les mots non voyellés peuvent être divisées en deux groupes : **les règles de support** et **les règles secondaires**. Il est important de noter que les règles de support sont appliquées dans une première phase. Ainsi, la plupart de ces règles sont à l'origine de la production des graphèmes et des phonèmes des voyelles longues, ce qui facilite l'application des règles secondaires dans une deuxième phase. Les règles secondaires sont à l'origine de la production des voyelles courtes, Tanween et des phonèmes de «Sukun».

Les règles de support sont appliquées dans l'ordre suivant : les règles de graphèmes et les règles lettres étrangères, les règles des voyelles longues, la règle de Ya-Maqsoura, la règle de gémination, les règles de ligature,les règles d'élision et les règles Shamsi et Ghamari.

Une fois les règles de support appliquées en se basant sur les trois principes mentionnés ci-dessus, les règles phonétiques restantes, qui sont les règles secondaires, sont appliquées dans l'ordre suivant : les règles des voyelles courtes et Sukun sont appliquées par la suite et les règles de Tanween sont appliquées en dernier lieu.

- ***Procédure de conversion Graphème en Phonème***

Dans ce qui suit nous allons expliciter la procédure de conversion G2P de quelque type de mot en s'appuyant sur des exemples bien précis. Nous débutons par le premier type de mot qui est constitué par des expressions contenant une voyelle longue. Par la suite, nous allons présenter le processus de phonétisation de chaque voyelle longue qui sont dans l'ordre suivant (les voyelles longues : ا /a :/ « Alif », ي /i :/ « Ya », و /u :/ « Waw » et وَا /u :/ « Waw et Alif »).

À priori, nous avons localisé dans le mot la position de ا /a :/ « Alif ». La règle de phonétisation de ا /a :/ « Alif » est appliquée. L'application de cette règle nous donne comme résultat le phonème correspondant à la voyelle longue ا qui est AE. S'il y a des autres possibilités pour appliquer une ou plusieurs autre(s) règles d'appui, on doit les appliquées. Une fois toutes les règles d'appui sont appliquées dans la première phase, cela nous donne un premier résultat de concaténation de phonèmes. Ensuite dans la deuxième phase et en se basant sur le principe que « chaque voyelle longue est toujours suivie et précédée par une consonne muette», les règles secondaires sont appliquées afin d'ajouter les phonèmes des voyelles courtes ou «Sukun». Ainsi, l'application des règles secondaire nous donne le résultat final qui est la phonétisation finale du mot qui peut avoir une ou plusieurs possibilités. Dans le tableau 5.1 un exemple illustratif est donné.

Mot		
Mot	**Règles & Explications**	**La phonétisation finale**
لَازم	**Phase 1 : application des règles d'appui**	L AE : Z AH M L AE : Z IH M L AE : Z UH M
	* Le graphème «ا» est une voyelle longue donc on obtient directement le phonème « AE : » qui lui correspond. * En appliquant les règles de graphèmes on obtient le phonème « L » qui correspond au graphème ل, le phonème « Z » qui correspond au graphème ز et le phonème « M » qui correspond au graphème م. L'application des règles d'appui nous donne comme résultat les phonèmes : L AE : Z M	
	Phase 2 : application des règles secondaires	
	En se basant sur le principe qu'un mot en dialecte tunisien se termine par une consonne muette (avec Sukun) ou une voyelle longue, on obtient directement le phonème « M » qui correspond au graphème م /m/ sans ajouter aucun phonème de voyelle courte. Dans ce cas on a deux consonnes successives, à la fin du mot, qui sont muettes « Z » et « M » alors on doit ajouter le phonème de voyelle courte entre ces deux consonnes qui sont : « AE » ou « IH » ou « UH » L'application des règles secondaires nous donne comme résultat les phonèmes : 1. L AE : Z AH M 2. L AE : Z IH M 3. L AE : Z UH M	

Table 5.1 – Les étapes de conversion G2P du mot لَازم

Rappelons que ي /i :/ « Ya », و /u :/ « Waw » peuvent être des voyelles longues dans certain mot et des consonnes dans d'autre mot. Pour cette raison, elles sont appelées des semi-consonnes. Dans la partie suivante, nous allons donner un exemple de conversion G2P de و dans les cas où elle est une consonne ou une voyelle longue.

- **و est une consonne.**

Nous allons pris le mot وقتَاش /wqtA$/ [quand] qui contient le « waw » comme un exemple illustratif pour présenter le processus de phonétisation.

Mot		
Mot	**Règles & Explications**	**La phonétisation finale**
وقتَاش	**Phase 1 : application des règles d'appui**	W AE Q T AE : SH W IH Q T AE : SH W UH Q T AE : SH
	* Le graphème ا est une voyelle longue donc on obtient directement le phonème « AE :» qui lui correspond. * Le graphème و dans cet exemple est une consonne car il se situe au début d'un mot. De ce fait, on obtient directement le phonème « W » qui lui correspond. * En appliquant les règles de graphèmes on obtient le phonème « Q » qui correspond au graphème ق, le phonème « T » qui correspond au graphème ت et le phonème « SH » qui correspond au graphème ش. L'application des règles d'appui nous donne comme résultat les phonèmes : W Q T AE : SH	
	Phase 2 : application des règles secondaires	
	* En se basant sur principe qu'un mot en Dialecte Tunisien se termine soit par une consonne muette ou voyelle longue, on obtient alors le phonème « SH » qui correspond au graphème ش. * En se basant sur le principe que chaque voyelle longue est toujours suivie et précédée par une consonne muette, dans ce cas le phonème de l'un des voyelles courtes doit être ajouté entre le premier et le deuxième phonème. L'application des règles secondaires nous donne comme résultat les phonèmes : 1. W AE Q T AE : SH 2. W IH Q T AE : SH 3. W UH Q T AE : SH	

Table 5.2 – Les étapes de phonètisation du mot وقتَاش /wqtA$/ [quand]

La lettre « waw » est une voyelle longue lorsqu'elle est une partie de وَاو الجمَاعة par exemple dans le mot خرجوَا /xrjwA/ [ils sortent], و + ا est une voyelle longue à la fin du mot.

Mot		
Mot	**Règles & Explications**	**La phonétisation finale**
خرجوَا	**Phase 1 : application des règles d'appui**	KH AE R JH UW KH IH R J UW KH UH R J UW
	* Le graphème «W» est une partie de وَاو الجمَاعة. Ainsi, on doit appliquer la règle de voyelle longue pour obtenir directement le phonème « UW ». * En appliquant les règles de graphèmes on obtient le phonème « KH » qui correspond au graphème خ, le phonème « R » qui correspond au graphème ر et le phonème « JH » qui correspond au graphème ج. L'application des règles d'appui nous donne comme résultat les phonèmes : KH R JH UW	
	Phase 2 : application des règles secondaires	
	* En se basant sur le principe que chaque voyelle longue est toujours suivie et précédée par une consonne muette, dans ce cas, l'insertion de phonème de l'un des voyelles courtes doit être seulement entre le premier et le deuxième phonème. => L'application des règles secondaires nous donne comme résultat les phonèmes : 1. KH AE R JH UW 2. KH IH R J UW 3. KH UH R J UW	

Table 5.3 – Les étapes de conversion G2P du mot خرجوَا /xrjwA/ [ils sortent]

Dans certain cas il n'y a aucune indication que le « waw » est une voyelle longue ou consonne. De ce fait, nous avons choisi de donner la phonétisation de ces deux possibilités. Dans le tableau suivant nous avons présenté le mot ثوم /vwm/ [l'ail] en tant qu'un exemple indicatif.

Mot		
Mot	**Règles & Explications**	**La phonétisation finale**
ثوم	**Phase 1 : application des règles d'appui**	TH UW M TH AE W M
	* Le graphème «W» est au sein d'un mot, alors il peut être une consonne ou une voyelle longue. Dans ce cas on doit appliquer en premier lieu la règle de voyelle longue pour obtenir directement le phonème « UW » dans une première possibilité. Ensuite on applique la règle de graphème pour obtenir le phonème « W » dans une deuxième possibilité. * En appliquant la règle de « Alif et lem » ال suivie par lettre Shamsi ث, on obtient la séquence de phonèmes «E IH TH TH ». * En appliquant les règles de graphèmes on obtient le phonème « TH » qui correspond au graphème ث, le phonème « M » qui correspond au graphème م. L'application des règles d'appui nous donne comme résultat les deux possibilités de phonèmes : 1. E IH TH TH UW M 2. E IH TH TH W M	
	Phase 2 : application des règles secondaires	
Possibilité 1 : E IH TH TH UW M	* En se basant sur le principe qu'un mot en Dialecte Tunisien se termine soit par une consonne muette (avec Sukun) ou voyelle longue, et dans que chaque voyelle longue est toujours suivie et précédée par une consonne muette, dans ce cas il est nécessaire de garder la liste des phonèmes de la première phase.	
Possibilité 2 : E IH TH TH W M	En se basant sur le principe qu'un mot en Dialecte Tunisien se termine soit par une consonne muette ou une voyelle longue, il faut garder le denier phonème « M » muette « sans phonème de voyelle courte ». Nous avons remarqué qu'en dialecte Tunisien si le mot est composé de trois consonnes et le « w » correspond à la deuxième consonne, la prononciation correspond au phonème de Fatha suivie par waw et cela se prononce comme « AE W ». Ainsi, on doit ajouter le phonème de Fatha « AE » avant le phonème « W».	

Table 5.4 – Les étapes de conversion G2P du mot ثوم /vwm/ [l'ail]

Maintenant nous passons à un autre type de phonétisation de mot qui contient une Shadda.

Mot		
Mot	**Règles & Explications**	**La phonétisation finale**
تفضّل	**Phase 1 : application des règles d'appui**	T F AE DD DD AE L
	* On applique la règle de gémination (shadda) qui fait doublement de consonne, dans ce cas le phonème « DH » qui correspond au graphème de ض va être doublé. * En se basant sur ce principe « un mot avec shadda, la consonne qui porte le shadda et celle qui la précède doivent porter aussi soit une voyelle courte soit une voyelle longue ». Dans notre cas, il n'a y pas de voyelle longue ainsi, on doit ajouter le phonème de voyelle courte avant et après le phonème répété « DH » « le phonème est répété à cause de Shadda ». En fait, on ajoute toutes les possibilités des voyelles courtes. * En appliquant les règles de graphèmes on obtient le phonème «T» qui correspond au graphème ت, le phonème «F» qui correspond au graphème ف et le phonème «L» qui correspond au graphème ل. L'application des règles d'appui nous donne comme résultat les phonèmes : 1. T F AE DH AE L 2. T F AE DH IH L	
	Phase 2 : application des règles secondaires	
	* En se basant sur le principe qu'on ne peut pas avoir deux consonnes successives qui portent une voyelle (longue ou courte) et le principe qu'un mot en Dialecte Tunisien se termine par une consonne muette, on doit garder le dernier phonème de «L» sans ajouter le phonème de voyelle courte. L'application des règles secondaires nous donne comme résultat les phonèmes 1. T F AE DH AE L 2. T F AE DH IH	

Table 5.5 – Les étapes de phonètisation de mot.

5.4 Evaluation

Dans cette section, nous présentons une évaluation de notre méthode de conversion graphème en phonème. Pour ce faire, nous avons utilisé un corpus comportant 7000 mots. En effet, nous avons créé trois types de corpus qui sont standardisés par CODA (L'orthographe conventionnel pour l'arabe Dialectal) : TARIC (Le Corpus d'interaction des chemins de fer tunisiens arabes). Nous avons recueilli des textes des blogs tunisiens de domaines différents (politique, le sport, la culture, la science ...). Aussi, nous avons utilisé notre outil qui nous permet de convertir le texte du dialecte tunisien écrit avec l'alphabet latin en caractères arabes suivant la convention de l'orthographe CODA de l'arabe dialectal.

Dans l'objectif d'évaluer les deux méthodes de phonétisation (avec et sans voyelles), nous avons réalisé une version voyellée du corpus, et vu l'absence d'un outil de voyellation automatique du corpus, celle-ci a été effectuée manuellement.

5.4.1 Présentation de l'outil d'évaluation

L'outil Sclite est créé par l'institut NIST (National Institue of Standards and Technology). Cet outil calcule le taux d'erreurs des mots (Word Error Rate : WER) dans notre cas l'outil génére le taux d'erreurs des phonèmes (Phoneme Error Rate : PER). En fait, ce taux d'erreur représente la mesure classique la plus utilisée pour l'évaluation des SRAP. Ainsi, cette mesure est obtenue en effectuant le calcul de la distance minimale entre la transcription du système et la référence créée manuellement par un expert linguistique.

Le processus d'évaluation de l'outil Scilte est réalisé en deux étapes. La première étape consiste à savoir « l'alignement textuel » dans lequel l'outil utilise un algorithme de programmation dynamique pour minimiser la distance de Levenshtein entre deux chaînes de texte (une référence et son hypothèse correspondante). La deuxième étape permet de savoir le « scoring ». En effet, après avoir aligné les chaînes de référence et d'hypothèse, les taux d'erreur sont calculés selon l'équation suivante :

$$PER = \frac{Sub + Del + Ins}{N} \tag{5.1}$$

Avec :

- **N** : Nombre de mots de la référence.
- **Substitutions (Sub)** : mot mal reconnu dans la transcription par rapport à la référence.

- **Insertions (Ins)** : mot supplémentaire dans la transcription par rapport à la référence.
- **Délétions (Del)** : mot non reconnu dans la transcription par rapport à la référence.

L'outil Sclite génère un rapport détaillant les résultats statistiques concernant les taux d'erreurs, les taux de réussite, les suppressions, les insertions et les substitutions.

5.4.2 Résultats obtenus

Le tableau 5.6 résume les résultats obtenus avec l'outil Sclite pour la conversion G2P du corpus avec et sans voyelles.

Table 5.6 – L'évaluation de la conversion G2P en termes de taux d'erreur en phonème

	Mots		**Phonèmes**	
	Voyellé	**Non-Voyellés**	**Voyellés**	**No-Voyellés**
Substitutions	0%	0.3%	0%	0.1%
Insertions	0%	0.3%	0%	0.2%
Délétions	0%	0.2%	0%	0.1%
Taux d'erreur en phonème	0%	0.8%	0%	0.4%

Conformément au tableau 5.6, le système de la conversion G2P des mots voyellés en dialecte tunisien est considéré comme performant puisqu'il produit un résultat de 100%. Néanmoins, le résultat de la conversion G2P des mots non voyellés est moins performant par rapport à celle des mots voyellés.

5.4.3 Discussion

Le score global de la méthode de la conversion G2P a dépassé les 99% de phonèmes corrects lorsque les mots du test sont voyellés et 99,6% de phonèmes corrects lorsque les mots du test sont non-voyellés. En effet, pour analyser les erreurs, une liste de mots erronés a été compilée. Certes, les différentes sources d'erreurs à savoir l'absence de voyelles courtes dans les mots, la présence des mots étrangers, les mots irréguliers et certains noms propres produisent de fausses prononciations.

De ce fait les erreurs peuvent être classées comme suit :Les erreurs relatives aux mots non-voyellés ont été classées en fonction de ce qui suit :

- L'existence de mots en dialecte tunisien qui ne suivent pas le principe d'emplacement mentionné précédemment à savoir un mot qui contient deux consonnes successives, engendre suite à son mauvais emplacement un problème d'insertion ou de suppression de la voyelle courte.

- Des erreurs sont dues à l'insertion des voyelles courtes, c'est-à-dire lors de la conversion de graphème en phonème de certains mots, notre méthode insère des voyelles courtes à un emplacement incorrect. À titre d'exemple, le mot تكلمي qui se prononce comme suit تكلمي possède une chaine phonétique générée par notre système /T AE K L M IY/ alors que la forme phonétique correcte qui doit être associée à ce mot est /T K AE L M IY/. Donc notre système a inséré le phonème AE qui correspond au « Fatha » dans l'emplacement inadéquat.
- Certaines erreurs sont dues à la substitution de certaines voyelles courtes par d'autres. Prenons à titre d'exemple le mot non voyellé الشرطة qui se prononce comme suit /E IH SH SH UH R T AE/. Il a comme une chaine phonétique générée par notre système /E IH SH SH AE R TT AE/ alors que la phonétisation correcte qui doit être associée à ce mot est /E IH SH SH UH R TT AE/. Donc, nous avons remarqué que nous avons inséré le phonème AE qui correspond au « Fatha » au lieu du phonème UH qui correspond au « Dhamma ».
- Lorsque certains mots ne gardent pas les règles de prononciation régulières et n'existent pas dans notre base d'exception, leurs prononciations seront incorrectes.
- L'existence des noms propres dans notre corpus peut engendrer des erreurs de phonétisation.
- D'autres erreurs sont relatives à la catégorie grammaticale, qui peut être un nom propre, un numéro ou tout autre élément lexical. En fait, certains noms propres tunisiens ont des prononciations qui sont différentes de celles de langue. Par exemple, le nom عزيز peut être prononcé de deux manières différentes /Aziz /ou /Ezayiz/. Dans d'autres cas, certains de ces noms propres arabes sont d'origine étrangère et disposent des prononciations différentes de la norme standard« MSA ». Par exemple, le nom "Midea" (Marque de climatiseur en français) est d'origine européenne, mais il est utilisé en dialecte tunisien et prononcé correctement par la plupart des gens en fonction de sa prononciation française. Néanmoins, notre système actuel le transcrit en tant que /M IH D YAE/, ce qui est évidemment erroné par rapport à sa prononciation en français.

5.5 La conversion G2P : approche probabiliste

Dans la première partie de ce chapitre, nous avons présenté notre approche à base de règles pour la conversion G2P du dialecte tunisien. Le principe de cette approche est d'utiliser des règles phonétiques et une base d'exceptions. Généralement, les systèmes de conversion G2P utilisant cette approche permettent de fournir des dictionnaires de prononciation de bonne qualité. Pour atteindre ce résultat, une bonne connaissance de la langue et de ses règles phonétiques est nécessaire. Cependant, ce type de système est toujours susceptible de commettre des erreurs en raison de la présence d'un mot exceptionnel qui n'est pas pris en considération par le concepteur des règles. Pour cela, une approche « guidée par les données » surgit. Cette approche est basée sur l'idée

d'avoir suffisamment d'exemples donnés pour la prédiction de la prononciation des mots qui n'ont pas été vus lors de l'apprentissage. Autrement dit, ce système serait capable de découvrir les règles phonétiques lors de l'apprentissage au moyen d'exemples de données. Dans cette perspective, notre proposition opère par l'utilisation d'une approche hybride qui combine une première conversion G2P purement symbolique avec une deuxième conversion purement probabiliste.

Pour illustrer cet aspect hybride, nous proposons d'utiliser en premier lieu les règles de prononciation du dialecte tunisien pour générer un dictionnaire phonétique. De peur que la deuxième méthode probabiliste apprenne des exemples erronés, ce dictionnaire doit être validé et corrigé manuellement par des experts car il serait possible de trouver des erreurs au niveau de la forme phonétique. À cet égard, nous présentons dans ce chapitre les détails de notre approche de la conversion G2P basée sur une méthode probabiliste : Conditional Random Fields (CRF). Notre choix de CRF est motivé par le fait qu'il aboutira à des résultats de l'état de l'art pour de multiples tâches de NLP, qui ont toutes un aspect commun, notamment les alignements monotones. De plus, CRF donne une prédiction à long terme et suppose des conditions d'indépendance de l'état détendu par rapport à HMM [Irrina, 2011].

Pour la conversion probabiliste, généralement la plupart des méthodes pour la conversion G2P nécessitent tout d'abord un alignement entre les graphèmes et les phonèmes du corpus d'apprentissage. Cet alignement s'avère utile du fait qu'il permet à cette méthode d'apprendre les probabilités des séquences de graphèmes qui engendrent des phonèmes particuliers tout en prenant en considération un certain contexte de graphèmes et phonèmes entourant. Ainsi, cette méthode doit connaître explicitement les phonèmes générés par tel ou tel graphème. Par conséquent, la qualité du résultat de la conversion G2P dépend fortement de cet alignement. Dans le même ordre d'idées, CRF nécessite un alignement de type 1-à-1 entre les graphèmes et les phonèmes. Cependant, ce type d'alignement n'est pas présenté dans le corpus initial. Théoriquement, il est fourni par un modèle externe.

Par ailleurs, l'apprentissage et la prédiction de la conversion G2P avec CRF se déroulent en deux étapes :

- la première étape consiste à effectuer l'alignement de type 1-à-1 de tous les mots du dictionnaire d'apprentissage en termes d'associations graphème-phonème.
- Dans la deuxième étape, les modèles de CRF sont formés à l'aide de l'ensemble de données de formation alignées. Enfin, ces modèles sont évalués à l'aide de données de test.

5.5.1 Etape d'alignement

Comme mentionné ci-dessus, CRF exige un alignement de type 1-à-1 entre les graphèmes et les phonèmes. Toutefois, en dialecte tunisien, la séquence de graphèmes et la

séquence de phonèmes de mots sont souvent de longueur différente, ce qui est incompatible avec la structure de CRF. La différence de longueur est due à plusieurs facteurs dont nous allons présenter quelques-uns dans ce qui suit :

- Habituellement, les textes écrits en dialecte tunisien ne sont pas voyellés. En effet, les voyelles courtes sont invisibles au niveau de graphèmes. Cependant, à l'oral, les phonèmes de ces voyelles courtes se prononcent naturellement. Par exemple pour le mot كتب, il peut être prononcé soit : كتب Ktib "écrit" ou كتب Ktob "des livres".
- Dans de nombreux exemples de mots en dialecte tunisien, il existe un double graphème représenté par un phonème simple comme le cas de Waw jame3a [le pluriel en arabe] composé de Waw et Alif" : وَا UW. Dans d'autres exemples : le « Alif et lem », représenté par double graphème, est prononcé par un phonème simple.
- De l'autre côté, deux phonèmes correspondant à un graphème : « Alif et lem » suivis d'une consonne solaire d'où un doublement de phonème de cette consonne solaire lors de la prononciation. Ainsi, on a une consonne dans la partie de graphèmes et deux phonèmes qui correspondent à cette consonne. Dans un autre exemple, nous relevons parfois l'ajout du phonème /Ai/ pour soutenir la prononciation de la première consonne silencieuse d'un mot. Autrement dit, l'alif prosthétique est rétabli devant les noms dont la première consonne est menacée de devenir quiescente (que l'on ne prononce pas).
- Dans une autre situation, nous pouvons saisir un autre problème : il s'agit de certains graphèmes quiescentes. Exemple le ta-marbouta à la fin du mot.
- Enfin, l'exemple de shadda qui a comme rôle de doubler le phonème d'une consonne.

Ces divers cas montrent qu'il y a une différence de longueur entre la séquence de graphèmes et la séquence de phonèmes de mots en dialecte tunisien. Afin de résoudre ce problème, le symbole "-" (aucune lettre, aucun phonème) est introduit dans l'alignement, sur le côté d'un graphème ou / et un phonème.

Dans le but de produire des alignements entre les graphèmes et les phonèmes de nos données comme elle exige CRF, nous avons utilisé deux modèles externes d'alignement sur lesquels CRF a été formée. Le premier consiste à utiliser un outil d'alignement nommé GIZA++, tandis que, le deuxième admet l'utilisation d'outil JMM « le modèle conjoint multigrammes ».

Dans les deux sections suivantes, nous allons présenté le scénario d'alignement suivant ces deux modèles.

5.5.1.1 Alignement basé sur GIZA++

Nous employons la boîte à outils GIZA++ pour obtenir les alignements graphème-phonème. Rappelons que les outils GIZA++ [Och et Ney, 2003] sont adoptés par la

plupart des systèmes de traduction automatique. L'entraînement est réalisé à partir d'un corpus de deux langues différentes, alignées par phrases et par mots.

En ce qui nous concerne, GIZA++ traite l'ensemble des mots comme langue source et l'ensemble des prononciations comme langue cible. En effet, l'apprentissage de la correspondance entre ces deux formes de graphèmes et de phonèmes est modélisé comme un problème de traduction automatique. Plus précisément, le mot en arabe est segmenté en caractères et entre lesquels il y a un espace, de même pour la forme phonétique de ce mot. Par conséquent, GIZA++ traite la totalité du mot comme une phrase alors que les consonnes et les phonèmes comme des mots. Dans notre cas, le graphème est considéré comme mot de langue source et le phonème est considéré comme un mot dans la langue cible. L'alignement entre les graphèmes et les phonèmes de toutes les données de formation est effectué suite à l'apprentissage de GIZA++.

Le format du résultat d'alignement avec GIZA++ est différent de celui admis par l'outil CRF. Pour cela, nous appliquerons un certain nombre de prétraitements pour extraire les associations entre un graphème et un phonème. À cet égard, devant chaque graphème on met son phonème. En cas d'absence de graphème ou de phonème, on se contente de mettre un epsilon. La figure suivante montre le résultat de pré-traitement effectué afin d'extraire les associations de type 1-1 entre un graphème et un phonème.

5.5.1.2 Alignement basé sur JMM

Pour une deuxième proposition d'alignement, nous employons un outil JMM qui est basé sur le concept d'un graphone $\boldsymbol{q}$ faisant une paire f_q de séquence de graphèmes et e_q une séquence de phonèmes. Ainsi, une séquence de graphones $q = (f_q, e_s)$ est générée pour chaque mot à partir de sa forme orthographique et sa prononciation. Pour ce faire, nous avons utilisé la notion de « 0-1 » graphones pour l'alignement, ce qui signifie qu'un ou zéro phonème est autorisé à être aligné avec un ou zéro graphème.

Dans nos expériences, nous avons choisi un modèle de 8 grammes pour effectuer alignement graphème-phonème. De plus, nous avons utilisé 0-1 graphones pour l'alignement, ce qui signifie que soit un ou zéro phonème est autorisé à être aligné avec un ou zéro graphème.

5.5.2 Etape expérimentale

5.5.2.1 Les mesures de performance

Afin d'évaluer la qualité des modèles de la conversion G2P sur des données textuelles en dialecte tunisien, la détermination du taux d'erreur de phonèmes (PER) et le taux d'erreur de mot (WER) s'impose comme mesures de performance. Ces mesures sont définies comme étant la distance de Levenshtein divisée par le nombre de phonèmes dans la prononciation de référence respectivement comme la fraction de mots contenant au moins une erreur.

5.5.3 Les résultats expérimentaux

Nous avons réalisé deux types d'expérimentations : une expérimentation avec une simple prononciation, une autre avec multiple prononciations.

5.5.3.1 Seule génération de prononciation par mot

Comme première expérience portant sur la génération de prononciation unique pour chaque mot, nous avons réalisé trois études en vue de déceler les différents effets sur la prédiction de CRF en dialecte tunisien. La première étude porte sur la modification de la taille d'ensemble de formation, dévelopement et test. La deuxième étude concerne les deux alignements effectués. La troisième étude se manifeste sur le changement des caractéristiques « unigramme » et « bigrammes ». Les caractéristiques « unigramme » prennent en compte seulement le phonème courant. Celles de « bigrammes » utilisent les phonèmes actuels et antérieurs. Ces trois études vont être mieux expliquées dans les résultats obtenus et résumés dans les tableaux suivants.

- **Influence de la modification la taille de l'ensemble de la formation, développement et test pour la prédiction CRF**

Lors de la première expérience nous avons testé l'influence de l'évolution de taille des données d'apprentissage sur la prédiction de CRF. Pour cela, nous définissons différentes tailles des données et nous avons divisé ces données d'apprentissage, de développement et de test tel que présenté dans le tableau 5.7. L'étiquette "5K" énoncée dans le tableau 5.7 équivaut à 5000 exemples de prononciation.

Table 5.7 – *Taille de formation, développement et test par ensemble : 1 à 5.*

	formation	Dev	Test
Ensemble 1 (5K)	3.75K	0.25K	1K
Ensemble 2 (7K)	5.25K	0.35K	1.4K
Ensemble 3 (8K)	6K	0.4K	1.6K
Ensemble 4 (10K)	7.5K	0.5K	2K
Ensemble 5 (18K)	13.5K	0.9K	3.6K

Afin de montrer la performance de CRF sur nos données dialectales, nous avons comparé notre approche par rapport à une autre approche de l'état de l'art « JMM ». De ce fait, nous avons effectué les expériences de G2P (Apprentissage et de test) sur les mêmes données en utilisant le système Sequitur G2P et le système CRF++. Les résultats obtenus sont rassemblés dans le tableau 5.8.

Table 5.8 – PER de la conversion G2P pour CRF, JMM et Phonetisaurus utilisant les ensembles de 1 à 5

	CRF	**JMM**	**Phonetisaurus**
Ensemble 1	22.87%	23.57%	28.46%
Ensemble 2	21.54%	22.13%	25.28%
Ensemble 3	20.74%	21.51%	23.21%
Ensemble 4	19.74%	20.41%	21.54%
Ensemble 5	14.31%	16.32%	17.83%

Pour l'ensemble 1, nous avons obtenu un taux d'erreur de 22,87% (PER) pour CRF, un taux d'erreur de 28,46% (PER) pour phonetisaurus. A partir de l'analyse de ce tableau, nous avons conclu que même en utilisant seulement la moitié de l'ensemble de la données « 5K », le système CRF est toujours performant. En comparant la performance de CRF et de Sequitur G2P, nous avons constaté que l'amélioration de CRF par rapport au Sequitur G2P est plus importante pour toutes les tailles de données. La meilleure performance est obtenue pour l'ensemble 5 en utilisant le système de CRF : 14.31% (PER). Ce résultat présente une amélioration de 2,01% en comparaison avec Sequitur G2P.

- **Effet de différents alignements de prédiction de CRF++**

Pour la seconde expérience, nous avons observé l'effet de l'alignement sur la performance du modèle de CRF. De ce fait, nous avons utilisé les deux modèles externes pour produire ces alignements. D'une part, nous avons employé la boîte à outils GIZA++ pour produire un premier exemple d'alignement. D'autre part, le modèle JMM est utilisé pour obtenir un autre exemple d'alignement. Pour chacun des alignements résultant, CRF est formé avec les mêmes caractéristiques. Ainsi, seulement l'influence de l'alignement doit être observée.

Le tableau 5.9 présente les résultats pour la prédiction CRF en utilisant l'alignement JMM et l'alignement de GIZA++. En comparant ces résultats, nous avons remarqué une légère diminution de la performance de l'alignement JMM 0,22% (PER) par rapport à l'alignement de GIZA++. Dans cette expérience, nous avons utilisé Ensemble 5 (18K) dans le tableau 5.9 .

Table 5.9 – Effet de différents alignements sur les deux tâches de conversion G2P

	PER% GIZA++ alignm	PER% JMM alignm	WER% GIZA++ alignm	WER% JMM alignm
CRF	14.31%	14.09%	21.35%	20.48%

- **Effet des caractéristiques unigramme et bigrammes**

Concernant la troisième expérience, nous avons étudié l'effet de la modification des caractéristiques « unigramme » et « bigrammes » sur la prédiction de CRF. Pour clarifier les deux notions, les caractéristiques « unigrammes » prennent en compte que le phonème courant tandis que des fonctionnalités de bigrammes utilisent les phonèmes actuels et antérieurs. Dans ces expériences, nous avons examiné aussi l'effet de différentes largeurs de contextes de graphèmes que les caractéristiques peuvent couvrir. Par exemple, (± 2) signifie que les caractéristiques peuvent couvrir deux graphèmes précédant et suivant par rapport à la position actuelle.

Les résultats du tableau suivant (étiquetés "unig" et "bigr") suggèrent qu'il vaut mieux utiliser les fonctionnalités de « bigrammes » que celles de « unigrammes ». De plus, pour un contexte de graphème (± 1), il représente une amélioration statistiquement significative.

Table 5.10 – Effet des caractéristiques unigrammes et bigrammes sur la conversion G2P

	PER% JMM alignm (unigr)	PER% JMM alignm (bigr)
CRF (±1)	14.36%	14.09%
CRF(±2)	14.51%	14.33%
CRF(±3)	15.07%	15.10%
CRF(±4)	15.34%	15.27%

En variant nos expériences, l'ensemble de fonctionnalités mis en œuvre (les caractéristiques « bigrammes », différentes tailles de contextes graphèmes dans le même modèle CRF (± 1 ± 2 ± 3 ± 4) et divers alignements), nous ont permis d'observer une influence sur la qualité de prédiction de G2P. Dans cette expérience, le contexte graphème (± 1) et la prédiction CRF en utilisant l'alignement JMM a obtenu le meilleur résultat.

5.5.3.2 Génération multiple de prononciation par mot

Lors des expériences précédentes, une seule prononciation par mot a été générée alors que dans les expériences suivantes, plusieurs prononciations pour un mot vont être générées. En effet, une «bonne» approche de conversion G2P devrait générer toutes les variantes de prononciation possibles pour un mot.

Dans cette section, nous étudions les n-meilleures sorties de G2P. Pour chaque mot, nous générons les n-meilleures prononciations de mots. En effet, nous avons évalué la qualité de n-meilleures pour n = 4. La mesure de performance utilisée est basée sur

les mesures de rappel et de précision. Le rappel (R) est le nombre de variantes de prononciation correcte généré divisé par le nombre total de variantes de prononciation de référence ; la précision (P) est le nombre de variantes de prononciation correcte divisé par le nombre total de variantes de prononciation générées.

A noter qu'un filtre (T) a été créé qui permet de ne pas prendre en considération des phonèmes d'hypothèse dont le pourcentage de prédiction est inférieur à (T=35%). L'élimination de ces phonèmes est due suite à l'impact négatif sur la prononciation du mot. C'est-à-dire que la prononciation du mot n'appartient pas au dialecte tunisien.

Le Rappel et la précision de la conversion G2P de CRF et de JMM sont présentés dans le tableau ci-dessous. Ainsi, le tableau 5.11 montre les résultats obtenus pour 4-meilleures prononciations de chaque mot avec différentes tailles des données d'apprentissage.

Table 5.11 – Rappel et précision pour la conversion G2P de CRF et JMM en utilisant les n-meilleures prononciations pour chaque mot

	CRF			JMM		
	Ens 3	Ens 4	Ens 5	Ens 3	Ens 4	Ens 5
R	88.51%	90.04%	91.41%	84.45%	85.75%	87.15%
P	84.45%	86.73%	87.13%	80.42%	82.93%	83.46%

Le meilleur résultat de n-meilleure prononciation est obtenu avec le modèle CRF en utilisant «Ens 5» : 91,41% de rappel et 87,13% de précision. La conversion G2P de CRF du dialecte tunisien donne un PER et un rappel élevé par rapport à d'autres langues telles que le français et l'anglais. Ceci est dû à l'absence de voyelles courtes dans l'apprentissage et le test corpus.

Pour conclure, afin de choisir la meilleure prédiction de CRF sur les données de dialecte tunisien, plusieurs expériences ont été réalisées. Dans cette optique, nous avons effectué plusieurs études. La première étude porte sur la modification de la taille d'ensemble de données d'apprentissage, de développement et de test. La deuxième étude concerne les deux alignements effectués. La troisième étude se manifeste sur le changement des caractéristiques « unigrammes » et « bigrammes ». De plus, nous avons testé l'impact d'une seule génération et de génération multiple de prononciation d'un mot.

D'après ces expériences, nous avons constaté que la meilleur prédiction avec CRF pour le cas de dialecte tunisien est d'utiliser JMM pour l'alignement, les caractéristiques « bigrammes » et CRF(+1,-1) comme taille de contextes graphèmes et utiliser une multiple génération d'un mot.

5.6 Conclusion

Dans le présent chapitre, nous avons essayé d'abord de cerner les principaux problèmes de conversion G2P du dialecte tunisien. Ces problèmes sont principalement dus

à la présence de certaines irrégularités dans le système orthographique de ce dialecte. Ils concernent l'ordre morpho-phonémiques, la liaison, l'élision, la présence de mots étrangers, l'apparition de nouveaux phénomènes comme l'assimilation et métathèses et les variations phonologiques et phonétiques. Ensuite, nous avons donné un survol des différentes solutions qui ont été suggérées pour résoudre ces problèmes et les règles utilisées pour la tâche de conversion G2P. A cette fin nous avons utilisé l'approche à bases de règles pour réaliser cette tâche.

Dans la deuxième partie, nous avons exposés les étapes de notre approche probabiliste en utilisant le modèle : CRF. Par ailleurs, l'apprentissage et la prédiction de la conversion G2P avec CRF se déroulent en deux étapes : ***i)*** la première étape consiste à effectuer l'alignement de type 1-à-1 de tous les mots du dictionnaire d'apprentissage en termes d'associations graphème-phonème. ***ii)*** Dans la deuxième étape, les modèles de CRF sont formés à l'aide de l'ensemble de données de formation alignées. Enfin, ces modèles sont évalués à l'aide de données de test.

Cependant, l'alignement entre les graphèmes et les phonèmes en dialecte tunisien est de type n-m ce qui est incompatible avec le besoin de CRF. Pour cette raison, nous avons utilisé deux modèles différents afin de générer des exemplaires de formation alignés : JMM et GIZA++, et ce en vue de vérifier lequel des deux modèles a le plus d'efficacité concernant le dialecte tunisien.

Au niveau d'apprentissage de CRF, nous avons réalisé deux types d'expérimentations : une expérimentation avec une simple prononciation, une autre avec de multiples prononciations. De plus et dans le but de choisir la meilleure prédiction de CRF sur les données dialectales, plusieurs expériences ont été réalisées. Plusieurs études ont été faites en vue de déceler les différents effets sur la prédiction de CRF en dialecte tunisien. La première étude porte sur la modification de la taille d'ensemble de formation. La deuxième étude concerne les deux alignements effectués. La troisième étude se manifeste sur le changement des caractéristiques « unigramme » et « bigrammes ».

Compte tenu de ce qui précède,dans le chapitre suivant nous allons présenter une vue d'ensemble sur le développement de notre premier SRAP pour le dialecte tunisien dans le domaine de renseignement ferroviaire et les résultats obtenus.

Chapitre 6

Premier SRAP pour le dialecte tunisien dans le domaine de renseignement ferroviaire

Sommaire

6.1 Introduction

L'évaluation de notre SRAP est une tâche déterminante pour mesurer son performance pour le dialecte tunisien. Dans lensemble de nos expériences menées dans ce chapitre, nous allons utiliser notre corpus TARIC. Nous allons utiliser le taux d'erreur de mots (Word Error Rate : WER) pour déterminer la performance de notre SRAP.

Ce chapitre retrace nos efforts pour le développement d'un premier SRAP pour le dialecte tunisien dans le domaine de renseignement ferroviaire. Certes, le SRAP peut être traité selon deux niveaux le premier est celui de l'apprentissage et le deuxième concerne le test et d'évaluation. De ce fait, nous allons commencer par présenter ces deux niveaux. Ensuite, nous décrivons les caractéristiques du corpus TARIC utilisées lors des expérimentations menées. Puis, nous détaillons les résultats dévaluation obtenus. Enfin, nous clôturons par une discussion afin dinterpréter les résultats obtenus.

6.2 Corpus d'apprentissage, développement et de test

Afin de définir les parties de notre corpus TARIC utilisées pour nos évaluations, nous avons divisé ce corpus en trois parties. La première partie du corpus est exploitée pour lapprentissage et elle représente de lordre de 90% de la taille totale, tandis que la deuxième partie constitue 5% du corpus utilisée pour le développement et la troisième partie est de l'ordre de 5% de taille total. Le tableau 6.1 représente les caractéristiques de ces deux différentes parties en termes de nombre d'heures, nombre d'énoncés et nombre de mots.

Table 6.1 – Caractéristiques du corpus dapprentissage, développement et de test

	# d'heures	**# d'énoncé**	**taille de vocabulaire**
Appr	8 Heures et 57 Minutes	18027	3027
Dev	33 Minutes and 40 Secondes	1052	612
Test	43 Minutes and 14 Secondes	2023	1009

Table 6.2 – Distribution de locuteurs en train, développement et test

	Nombre de locuteur		**# Locuteur non vus (temps)**
	Mâle	**Femelle**	
Train	60	37	-
Dev	3	8	7(15min 19Sec)
Test	1	5	4(24 min 15 Sec)

Le corpus TARIC comprend plus de 21k énoncés dont 108 locuteurs différents (certains locuteurs sont partagés sur apprentissage, dev et test). Le corpus contient 82K mots en cours d'exécution avec un vocabulaire de 3207 mots. Comme il est présenté dans le tableau 6.2, environ 9,5 heures des données transcrites sont consacrées à l'apprentissage et le développement, tandis que les 43 minutes restantes composent l'ensemble de test.

6.3 Modèle acoustique

L'évaluation de la performance d'un SRAP du dialecte tunisien est faite par le biais des expérimenta-tions menées avec KALDI. Notre modèle acoustique est formé en utilisant des caractéristiques PLP (Perceptual linéaire prédictive). De ce fait, pour chaque frame, nous incluons aussi ses voisins ±4 frame et appliquer transformation linéaire discriminante Analyse (LDA) afin de projeter les "frames" concaténés à 40 dimensions, suivie par Maximum Likelihood Linear Transformer (MLLT). Nous avons également appliqué une adaptation du locuteur à travers la technique de caractéristique de de l'espace Maximum Likelihood Régression Linéaire (fMLLR).

Par ailleurs, ces modèles sont tous les modèles standards de 3 états dépendant du contexte et triphones. Le modèle GMM-HMM a environ 15K gaussiennes pour 2.5K états liés.

6.4 Modèle du langage

En ce qui concerne la modélisation du langage (LM), elle a été construite en utilisant les transcriptions de TARIC décrits dans le chapitre 4. Un corpus total de 100K mots est utilisé. Etant donné le manque de norme d'écriture pour le dialecte tunisien, ce corpus est normalisé en utilisant la convention d'écriture orthographique CODA. Nous avons entraîné différents modèles de langage 3-grammes en utilisant la boîte à outils SRILM. De même, le modèle de langage est fait avec modification de la méthode de lissage Kneser-Ney modifiée [Kneser Ney, 1995]. Ces modèles ont été interpolées pour créer la LM finale en minimisant la perplexité (ppl) sur le corpus de développement. Compte tenu de la taille limitée de nos données de formation, nous n'avons pas appliqué des seuils sur le lexique, ni la taille du modèle de langage final.

6.5 Expérimentations

Le système KALDI produit les treillis comme résultat de la reconnaissance. Le modèle de langage est optimisé sur l'ensemble de développement et utilisé pour calculer le meilleur chemin à travers le treillis.

Pour l'évaluation de notre SRAP, nous avons utilisé l'outil Sclite. Rappelons que cet outil calcule le taux d'erreur de mots (Word Error Rate : WER). En fait, ce taux d'erreur représente la mesure classique la plus utilisée pour l'évaluation des SRAP. En sus de ce taux, l'outil permet de générer un rapport détaillant les résultats statistiques concernant les suppressions, les insertions et les substitutions.

Table 6.3 – Résultats d'évaluation de la première SRAP pour le dialecte tunisien

	LM PPL	**WER (%)**	**Substitutions**	**Délétion**	**Insertions**
Dev	41.69	21.5	15.2	4.1	2.3
Test	53.71	22.6	16.0	4.0	2.5

Les résultats d'évaluation de SRAP pour le dialecte tunisien sont présentés dans le tableau 6.3. L'évaluation de notre système donne lieu à un taux WER de 22,6% sur l'ensemble de test. En examinant les résultats, nous avons constaté que les 6 premiers substitutions fréquentes sont des mots avec Shadda qui représentent le doublement des consonnes.

Ces erreurs seront profondément étudiées dans les travaux futurs en vue d'améliorer notre SRAP pour le dialecte tunisien.

6.6 Conclusion

Nous avons présenté dans ce chapitre nos efforts menés pour développer et évaluer notre premier SRAP pour le dialecte tunisien dans le domaine de renseignements ferroviaire. Pour ce faire, nous avons utilisé le corpus TRAIC pour l'apprentissage et le test de notre SRAP. Ce corpus est normalisé en utilisant la convention d'écriture orthographique CODA.

Nos expériences sont faites selon deux niveaux à savoir, l'apprentissage et le test. L'évaluation donne un taux d'erreur de mots de 22.6%.

Conclusion et perspectives

Les travaux menés dans cette thèse s'intègrent dans le cadre des recherches sur les systèmes de reconnaissance automatique de la parole. Cette problématique a été abordée pour le dialecte tunisien dans un domaine limité à savoir les renseignements ferroviaires qui constitue un domaine cerné en termes de vocabulaire utilisé.

Afin d'approcher ce domaine de recherche, nous avons commencé par la présentation des caractéristiques de la langue arabe en passant du littéral vers le dialectal. Nous avons choisi le dialecte tunisien comme étant un exemple de l'arabe dialectal. Ensuite, nous avons passé à la définition des concepts de base inhérents de la reconnaissance de la parole. Par la suite, nous avons évoqué un aperçu sur quelques SRAP développés dans la littérature pour les langues peu dotées. En deuxième lieu, nous avons détaillé les approches pour la conversion graphème en phonème proposées dans la littérature.

La présence d'un corpus est primordiale. L'élaboration du corpus TARIC représente le point de départ dans cette thèse. Il s'agit d'une ressource qui regroupe des audio et ses transcriptions en dialecte tunisien relatifs à des renseignements ferroviaires enregistrés à la gare de Tunis de la SNCFT. À travers ce corpus, nous avons effectué en premier lieu une étude linguistique qui porte sur différents niveaux à savoir le niveau phonologique, morphologique, syntaxique et lexical. En deuxième lieu, nous avons dégagé des règles phonétiques qui vont servir après à la conversion G2P.

Vu l'absence des normes de transcription orthographique pour le dialecte tunisien, nous avons proposé un guide qui regroupe un ensemble de règles de transcription dans la mesure d'avoir une orthographe bien définie et bien harmonisée. Ainsi, ce guide nommé CODA constitue une adaptation du travail de [Habash 2012] du dialecte égyptien vers le dialecte tunisien. Compte tenu des lignes directives de CODA, nous avons transcrit notre corpus TARIC.

Nous avons ensuite détaillé notre méthode pour la conversion graphème en phonème pour le dialecte tunisien. L'originalité de cette méthode réside à combiner une méthode à base de règles avec une méthode probabiliste en utilisant les modèles CRF (Conditional Random Fields). Pour illustrer cet aspect hybride, nous proposons d'utiliser en premier lieu les règles de prononciation et un lexique d'exceptions. Le rôle de cette conversion est de produire un dictionnaire de prononciation. Afin que la deuxième méthode probabiliste

n'apprenne pas des exemples erronés, une validation par un expert est nécessaire. En deuxième lieu, ce dictionnaire est considéré de ce fait comme une entrée pour l'approche probabiliste : CRF. Cependant, CRF nécessite un alignement de type 1-à-1 entre les graphèmes et les phonèmes. Généralement, ce type d'alignement n'est pas présenté dans le corpus initial. Dans ce travail, nous avons utilisé les deux modèles JMM et GIZA pour effectuer l'alignement sur lequel une CRF a été formée.

L'hybridation permet de tirer profit des avantages de l'approche symbolique et des modèles CRF. L'apport de ces modèles réside dans la prise en compte de la corrélation forte entre les graphèmes et les phonèmes d'un même mot afin de prédire une conversion G2P. Par ailleurs, l'apport de l'approche symbolique a permis de surmonter les insuffisances des modèles CRF qui nécessitent un corpus annoté. Aussi, ce type d'approche permet d'intégrer des connaissances phonétiques pour améliorer cette conversion G2P. Les résultats de l'évaluation des différentes méthodes proposées ont mis en relief les performances de la méthode hybride.

Afin d'évaluer notre méthode hybride, un corpus dialectal volumineux est nécessaire. De ce fait, nous nous sommes intéressés à la création de corpus pour le dialecte tunisien. Le défi consiste à surmonter les problèmes de la carence des données. Ainsi, nous avons essentiellement collecté des ressources textuelles issues à partir de deux méthodes qui sont l'aspiration des blogs et la translittération des données dialectales écrites en caractère latin vers les caractères arabes.

À l'issue de ces traitements, ces ressources notamment TARIC et l'outil de conversion G2P sont utilisées pour développer le premier SRAP pour le dialecte tunisien dans le domaine de renseignement ferroviaire. Les résultats de ce système ont donné un taux de 22,6% de WER.

Dans la continuité de notre thèse, nous pouvons distinguer plusieurs perspectives de recherche. À court terme, des améliorations pourraient être apportées à la conversion G2P pour les mots non voyellés. En effet, afin d'améliorer la robustesse de cet outil face aux erreurs produites à cause de l'absence des signes diacritiques, nous pensons qu'il serait utile d'utiliser un outil de voyellation automatique de dialecte tunisien. Cela permet d'ajouter automatiquement les signes diacritiques et par conséquence la détermination de la forme phonétique devient plus facile.

Aussi, nous envisageons d'utiliser d'autres fonctionnalités dans la modélisation de CRF telles que l'annotation de POS du dialecte tunisien. Cela permet d'améliorer la performance de la prédiction de la conversion G2P. Les travaux récents sur ce type d'annotation du dialecte tunisien, au sein de notre laboratoire, sont en faveur de l'application concrète de cette perspective [Boujelbane 2014].

Comme perspectives à moyen terme, nous souhaitons exploiter le modèle HCRF (Hidden Conditional Random Fields) pour la conversion G2P. Ce modèle traite les correspondances arbitraires entre les chaînes de graphèmes et phonèmes observées dans les

données d'apprentissage. Autrement dit, nous n'avons pas besoin d'un module externe pour faire l'alignement 1-à-1 entre les graphèmes et les phonèmes.

Le corpus exploité dans cette étude est un corpus transcrit manuellement vu l'absence des outils de transcription automatique pour le dialecte tunisien. De ce fait, nous pouvons tirer profit de notre premier SRAP pour le dialecte tunisien afin de transcrire d'une manière semi-supervisée d'autres audio enregistrés dans les guichets de la gare. Il s'agit, de corriger et valider seulement ses transcriptions.

À plus long terme, nous espérons pouvoir adapter notre SRAP du domaine de renseignement ferroviaire à un autre domaine le «news ». En fait, l'adaptation de ce système au domaine de «news » n'est pas considéré comme un passage de thème à un autre mais le passage d'un type de dialecte à un autre. Autrement dit, il s'agit de passage du dialecte familier au dialecte intellectualisé. Ainsi, le premier type de dialecte contient une masse importante de mots purement dialectaux. Cependant, le dialecte intellectualisé se présente comme un mélange entre le MSA et le dialecte tunisien.

Liste des publications

1. **Titre** : "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition"
   Auteurs : Abir Masmoudi, Mariem Ellouze, Yannick Estève, Lamia Hadrich Belguith and Nizar Habash
   Conférence : LREC'2014, Reykjavik, Iceland, 26-31 Mai 2014.
2. **Titre** : "A conventionnal Orthography for Tunisian Arabic"
   Auteurs : Ines Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith and Nizar Habash
   Conférence : LREC'2014, Reykjavik, Iceland, 26-31 Mai 2014.
3. **Titre** : "Phonetic tool for the Tunisian Arabic"
   Auteurs : Abir Masmoudi, Yannick Estève, Mariem Ellouze Khmekhem, Fethi Bougares and Lamia Hadrich Belguith
   Conférence : SLTU'2014, The 4th International Workshop on spoken Language Technologies for Under-resourced Languages, Russia, 14-16 Mai 2014.
4. **Titre** : "Phonétisation automatique du Dialecte Tunisien "
   Auteurs : Abir Masmoudi, Mariem Ellouze Khmekhem, Yannick Estève, Fethi Bougares, Sawssan dabbar and Lamia Hadrich Belguith.
   Conférence : JEP'2014, 30 éme Journée d'études sur la parole, Le Mans-France, 23-24 Juin 2014.
5. **Titre** : "Arabic Transliteration of Romanized Tunisian Dialect Text : A Preliminary Investigation"
   Auteurs : Abir Masmoudi, Nizar Habash, Mariem Ellouze, Yannick Estève et Lamia Hadrich Belguith
   Conférence : CICLing'2015, Egypt, 14-20 Avril 2015.
6. **Titre** : "Conditional Random Fields for the Tunisian Dialect Grpaheme-to-Phoneme conversion"
   Auteurs : Abir Masmoudi, Mariem Ellouze, Fethi Bougares, Yannick Estève and Lamia Belguith.
   Conférence : INTERSPEECH'2016, San Francisco, 9-12 Septembre 2016.

Bibliographie

[Al-Badrashiny 2014] Al-Badrashiny M., Eskander R., Habash N. et Rambow O., Automatic transliteration of romanized dialectal arabic, dans *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014.

[Alghamdi 2002] Alghamdi M., Elshafei M. et Husni-Al-Muhtaseb, Speech units for arabic text-to-speech, *Fourth Workshop on Computer and Information Sciences*, pages 199–212, 2002.

[Ammar 2012] Ammar Z., Larabe standard lu par des locuteurs tunisiens et des locuteursmarocains : Production des voyelles et des fricatives interdentales, 2012.

[Andersen 1996] Andersen O., Kuhn R., Lazaridès A., Dalsgaard P., Haas J. et Nth E., Comparison of two tree-structured approaches for grapheme-to-phoneme conversion, *Spoken Language Processing*, 3 :1700 1703, 1996.

[Asker 2009] Asker L., Argaw A. A., Gambäck B., Asfeha S. E. et Habte L. N., Classifying amharic web news, *Information Retrieval*, 12 :416–435, 2009.

[Baccouche 1974] Baccouche T., Esquisse dune étude comparative des schémas des verbes en arabe classique et en arabe tunisien, *Les cahiers de Tunisie*, 22 :87–88, 1974.

[Baccouche 1994] Baccouche T., L'emprunt en arabe moderne, *Beit Elhikma et IBLV*, 1994.

[Baccouche 2004] Baccouche T., Dialectes et dialectologie en linguistique arabe, *larabe dialectal : enqus, descriptions, interprtions, Colloque AIDA6*, pages 15–26, 2004.

[Baccouche 2000] Baccouche T. et Mejri S., L'atlas linguistique de tunisie : spécificités phonologiques, *Revue tunisienne des sciences sociales*, Numéro spécial : Langage et altté : l'expérience de l'atlas linguistique de Tunisie :157–162, 2000.

[Bahou 2010] Bahou Y., Masmoudi A. et Belguith L. H., Traitement des disfluences dans le cadre de la compréhension automatique de loral arabe spontané, *TALN'2010*, 2010.

[Béchet 2001] Béchet F., Lia_phon, un syst{eme complet de phonétisation de texte., *Traitement Automatique des Langues*, 42, 2001.

[Belgacem 2009] Belgacem M., Construction dun corpus robuste de diffnts dialectes arabes, *Actes des VIIIs RJC Parole*, 2009.

[Belguith 1999] Belguith L. H., *Traitement des erreurs daccord de larabe basr une analyse syntagmatique ndue pour la vfication et une analyse multicrit pour la correction.*, Thèse de doctorat, Facults Sciences de Tunis, 1999.

[Belguith 2006] Belguith L. H. et Chaaben N., Analyse et désambigution morphologiques de textes arabes non voyellés, *Traitement Automatique des Langues Naturelles (TALN2006)*, page 493501, 2006.

[BenFrah 2008] BenFrah A., Les affriquées en dialectal tunisien le phénomène de la tachtacha comme exemple, *L'Atlas tunisien*, 2008.

[Berment 2004] Berment V., *Méthodes pour informatiser des langues et des groupes de langues peu dotées*, Thèse de doctorat, Université Joseph Fourier Grenoble„ 2004.

[Besacier 2005] Besacier L., Le V. B., Castelli E., Sethserey S. et Protin L., Reconnaissance automatique de la parole pour des langues peu dotées : Application au vietnamien et au khmer, *TALN'2005*, 2005.

[Biadsy 2009] Biadsy F., Habash N. et Hirschberg J., *Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules*, page 397405, Annual Conference of the North American Chapter of the ACL, 2009.

[Billa 2002] Billa J., Noamany M., Srivastava A., Daben Liu R. S., J. Xu J. M. et Kubala F., Audio indexing of arabic broadcast news, *In Proceedings of ICASSP'2002*, page 58, 2002.

[Bilmes 2003] Bilmes J. et Kirchhoff K., *Factored Language Models and Generalized Parallel Backo*, chapitre Human Language Technologies North American Chapter of the Association for Computational Linguistics (HLT/NAACL)„ pages 4–6, 2003.

[Bisani 2008] Bisani M. et Ney H., Joint-sequence models for grapheme-to-phoneme conversion, *Speech Communication*, 50 :434–451, 2008.

[Bonafonte 1996] Bonafonte A. et Mariño J. B., Language modeling using x-grams, dans *The 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, 1996*, 1996.

[Bougares 2012] Bougares F., *Attelage de systèmes de transcription automatique de la parole*, Thèse de doctorat, Université du maine, 2012.

[Boujelbane 2014] Boujelbane R., Ellouze M., Béchet F. et Belguith L. H., De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens, *TAL*, 55(2) :73–96, 2014.

[Boukadida 2008] Boukadida N., *Connaissances phonologiques et morphologiques dvationnelles et apprentissage de la lecture en arabe (Etude longitudinale)*, Thèse de doctorat, Université Rennes 2, 2008.

[Brown 1992] Brown P., deSouza P., Mercer R., Pietra V. D. et Lai J., Class-based n-gram models of natural language, *Computational Linguistics*, 18 :467479, 1992.

[Brun 2007] Brun A., Langlois D. et Smaïli K., Improving language models by using distant information, *International Symposium on Signal Processing and its Applications (ISSPA)*, pages 1–4, 2007.

[Chelba 1997] Chelba C., Engle D., Jelinek F., Jimenez V., Khudanpur S., Mangu L., Printz H., Ristad E., Rosenfeld R., Stolcke A. et Wu D., Structure and performance of a dependency language model, *the European Conférence on Speech Communication and Technology (Eurospeech)*, pages 2775–2778., 1997.

[Chelba 2000] Chelba C. et Jelinek F., Structured language modeling, *Computer Speech and Language*, 14 :283–332, 2000.

[Church 2003] Church K., Speech and language processing : Where have been and where are we going ?, *Eurospeech2003*, 2003.

[Clarkson 1999] Clarkson P. et Robinson T., Towards improved language model evaluation measures, *the 6th European Conférence on Speech Communication and Technology (Eurospeech)*, 5 :1927–1930, 1999.

[Daelemans 1994] Daelemans W. et van den Bosch A., *A language-independent, data-oriented architecture for grapheme-to-phoneme conversion*, pages 199–203, 1994, URL `http://www.cnts.ua.ac.be/papers/1994/db94.pdf`.

[Daelemans 1996] Daelemans W. et van den Bosch A., Language-independent data-oriented grapheme-to-phoneme conversion, *In : Van Santen, J. P. H., Sproat, R. W., Olive, J. P., Hirschberg, J. (Eds.), . Springer.In Progress in Speech Synthesis*, page 77 90, 1996.

[Damper 1997] Damper R. I. et Eastmond J. F. G., Pronunciation by analogy : Impact of implementational choices on performance, *Language and Speech*, 40 :123, 1997.

[Dardour 2008] Dardour F., *Langue enseignée et dialecte Arabe : Quelle méthodoligie et quelle formation pour l'acquisition de la compétence communicative en arabe standard ?*, Thèse de doctorat, Universitncy 2, 2008.

[Davis 1990] Davis S. B. et Mermelstein P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *In Alex Waibel et Kai-Fu Lee, editeurs, Readings in speech recognition*, page 6574, 1990.

[Dedina 1991] Dedina M. J. et Nusbaum H. C., Pronounce : A program for pronunciation by analogy, *Computer Speech and Language*, 5 :55 63, 1991.

[DO 2006] DO T. N. D., *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*, Thèse de doctorat, Université de Grenoble, France, 2006.

[Dong 2011] Dong W. et Simon K., Letter-to-sound pronunciation prediction using conditional random fields, *Signal Processing Letters*, 18, 2011, URL `http://www.eurecom.fr/publication/3303`.

[Edgar 1967] Edgar P., *Swahili Language Handbook*, 1967.

[El-Imam 2004] El-Imam Y. A., Phonetization of arabic : rules and algorithms, *Computer Speech and Language*, 18 :339373, 2004.

[Elmahdy 2010] Elmahdy M., Gruhn R., Minker W. et Abdennadher S., Cross-lingual acoustic modeling for dialectal arabic speech recognition, *International Conference on Speech and Language Processing (Interspeech)*, 2010.

[Elmahdy 2014] Elmahdy M., Hasegawa-Johnson M. et Mustafawi E., Development of a tv broadcasts speech recognition system for qatari arabic, *The 9th edition of the Language Resources and Evaluation Conference : LREC 2014*, 2014.

[Elmahdy 2011a] Elmahdy M., Hasegawa-Johnson M., Mustafawi E., Duwairi R. et Minker W., Challenges and techniques for dialectal arabic speech recognition and machine translation, dans *Qatar Foundation Annual Research Forum*, 2011a.

[Elmahdy 2011b] Elmahdy M., Rainer G., Slim A. et Wolfgang M., Rapid phonetic transcription using everyday life natural chat alphabet orthography for dialectal arabic speech recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011b.

[Elmahdy 2012] Elmahdy M., Rainer G. et Wolfgang M., *Novel Techniques for Dialectal Arabic Speech Recognition*, Springer, Boston (USA), 2012.

[Elshafei 2006] Elshafei M., Al-Muhtaseb H. et Alghamdi M., Statistical methods for automatic diacritization of arabic text, *The Saudi 18th National Computer Conference*, 18 :301–306, 2006.

[Embarki 2008] Embarki M., Les dialectes arabes modernes : état et nouvelles perspectives pour la classification géo-sociologique, *Arabica, Brill Academic Publishers*, 55 :583–604, 2008.

[Eskander 2013] Eskander R., Habash N., Rambow O. et Tomeh N., Processing spontaneous orthography, *In Proceedings of Conference of the North American Association for Computational Linguistics (NAACL)*, 2013.

[Fishman 1967] Fishman J. A., Bilingualism with and without diglossia ; diglossia with and without bilingualism, *Social issues*, 23 :29–38, 1967.

[Franz 2003] Franz J. O. et Ney H., A systematic comparison of various statistical alignment models, *COMPUTATIONAL LINGUISTICS*, 29, 2003.

[Frédéric Bimbot 1995] Frédéric Bimbot E. L. e. B. A. Roberto Pieraccini, Variable-length sequence modeling : Multigrams, *Signal Processing Letters, IEEE*, pages 111–113, 1995.

[Gelas 2012] Gelas H., Abate S. T., Besacier L. et Pellegrino F., Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées á morphologie riche, *JEP-TALN-RECITAL 2012, Atelier TALAf 2012 : Traitement Automatique des Langues Africaines*, page 5362, 2012.

[Gibson 1998] Gibson M., *Dialect contact in Tunisian Arabic : Sociolinguistic and structural aspects*, Thèse de doctorat, University of Reading, 1998.

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

[Habash 2012] Habash N., Diab M. et Rambow O., Conventional orthography for dialectal arabic, dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718, 2012.

[Habash 2006] Habash N. et Rambow O., MAGEAD : A morphological analyzer and generator for the arabic dialects, dans *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006.

[Habash 2013] Habash N., Roth R., Rambow O., Eskander R. et Tomeh N., Morphological analysis and disambiguation for dialectal arabic, dans Vanderwende L., III H. D. et Kirchhoff K., rédacteurs, *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics*, pages 426–432, The Association for Computational Linguistics, 2013.

[Häkkinen 2003] Häkkinen J., Suontausta J., Riis S. et Jensen K. J., Assessing text-to-phoneme mapping strategies in speaker independent isolated word recognition, *Speech Communication*, 41(2-3) :455–467, 2003.

[Harrat 2015] Harrat S., Meftouh K., Abbas M., Jamoussi S., Saad M. et Smaïli K., Cross-dialectal arabic processing, *In Computational Linguistics and Intelligent Text Processing*, pages 620–632, 2015.

[Harrat 2014] Harrat S., Meftouh K., Abbas M. et Smaïli K., Grapheme to phoneme conversion-an arabic dialect case, *In Spoken Language Technologies for Under-resourced Languages : SLTU'2014*, 2014.

[Hermansky 1991] Hermansky H. et Jr L. A. C., Perceptual linear predictive (plp) analysis-resynthesis technique, *Eurospeech*, page 037 038, 1991.

[Holes 2004] Holes C., Modern arabic : Structures, functions and varieties, *Georgetown University Press*, 2004.

[Illina 2011] Illina I., Fohr D. et Jouvet D., Grapheme-to-phoneme conversion using conditional random fields, *Interspeech' 2011*, 2011.

[cal Imed jdouben 2012] cal Imed jdouben F. et Houacine A., Outil de transcription phonétique à partir du texte arabe, *Le 11ème Colloque Africain sur la Recherche en Informatique et en Mathàmatiques Appliquàes - CARI2012*, 2012.

[Jamoussi 2004] Jamoussi S., *Mèthodes statistiques pour la compréhension automatique de la parole*, Thèse de doctorat, L'université Henri Poincarancy1, 2004.

[Jensen 2000] Jensen J. et Riis S., Self-organizing letter code-book for text-to-phoneme neural network model, *Spoken Language Processing*, 3 :318 321, 2000.

[Juang 1992] Juang B.-H. et Rabiner L., Issues in using hidden markov models for speech recognition, *Advances in Speech Signal Processing*, pages 509–533, 1992.

[Kheang 2011] Kheang S., Iribe Y. et Nitta T., Letter-to-phoneme conversion based on two-stage neural network focusing on letter and phoneme contexts, *INTERSPEECH'2011*, pages 1885–1888, 2011.

[Kirchhoff 2005] Kirchhoff K. et Yang M., Improved language modeling for statistical machine translation, *the ACL Workshop on Building and Using Parallel Texts (ParaTex) Association for Computational Linguistics*, pages 125–128, 2005.

[Kneser 1995] Kneser R. et Ney H., Improved backing-off for n-gram language modeling, *ICASSP'1995*, 1995.

[Lajmi 2009] Lajmi D., Spécificités du dialecte sfaxien, *Synergies Tunisie numéro 1*, pages 135–142, 2009.

[Langlois 2000] Langlois D., Smaïli K. et Haton J. P., Dealing with distant relationships in natural language modelling for automatic speech recognition, *the World MultiConf. on Systemics, Cybernetics and Informatics (SCI)*, 6, 2000.

[Lars 2007] Lars A., Argaw A. A., Gambäck B. et Sahlgren M., Applying machine learning to amharic text classification, *5th World Congress of African Linguistics*, 2007.

[Laurent 2010] Laurent A., *Auto-adaptation et reconnaissance automatique de la parole*, Thèse de doctorat, Universit é du Maine, 2010.

[Laurent 2014] Laurent A., Meignier S. et Deléglise P. P., Improving recognition of proper nouns in asr through generating and filtering phonetic transcriptions, *Computer Speech and Language*, 28 :979996, 2014.

[Le 2006] Le V. B., *Reconnaissance automatique de la parole pour des langues peu dotées*, Thèse de doctorat, Université Joseph Fourier, 2006.

[Lecovré 2010] Lecovré G., *Adaptation thématique non supervisée dun système de reconnaissance automatique de la parole*, Thèse de doctorat, Université européenne de Bretagne, 2010.

[Lehnen 2011] Lehnen P., Stefan H., Andreas G. et Hermann N., Incorporating alignments into conditional random fields for grapheme to phoneme conversion, pages 4916–4919, Prague, Czech Republic, 2011.

[Loots 2011] Loots L. et Niesler T., Automatic conversion between pronunciations of different english accents, *Speech Communication*, 53 :7584, 2011.

[Maamouri 2004] Maamouri M., Buckwalter T. et Cieri C., Dialectal arabic telephone speech corpus : Principles, tool design and transcription conventions, *In NEMLAR International Conference on Arabic Language Resources and Tools, Cairo*, pages 22–23, 2004.

[Marchand 2000] Marchand Y. et Damper R. I., A multistrategy approach to improving pronunciation by analogy, *Computational Linguistics*, 26 :195219, 2000.

[Markel 1982] Markel J. et Gray A., Linear prediction of speech, *SpringerVerlag New York*, 1982.

[Mcculloch 1987] Mcculloch N., Bedworth M. et Bridle J., Netspeak ? a reimplementation of nettalk, *Computer Speech and Language*, page 289 302., 1987.

[Mejri 2003] Mejri S. et Baccouche T., Latlas linguistique de tunisie : repéres méthodologiques pour la description du systéme dialectal, *n : Lentin, J., Lonnet, A. (eds.) Mnges David Cohen,*, page 4754, 2003.

[Mejri 2009] Mejri S., Said M. et Sfar I., Pluringuisme et diglossie en tunisie, *Synergies Tunisie*, 1 :53–74, 2009.

[Niesler 1996] Niesler T. et Woodland P. C., A variable-length category-based n-gram language model, dans *IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings : ICASSP'1996*, pages 164–167, 1996.

[Nimaan 2006] Nimaan A., Nocera P. et Bonastre J.-F., Reconnaissance automatique de la parole en langue somalienne, *Les journées d'étude de la parole, JEP'2006*, 2006.

[Ouerhani 2009] Ouerhani B., Interférence entre le dialectal et le littéral en tunisie : Le cas de la morphologie verbale, *Synergies Tunisie nunm'ero 1*, pages 75–84, 2009.

[Pagel 1998] Pagel V., Lenzo K. et Black A. W., Letter-to-sound rules for accented lexicon compression, *Spoken Language Processing*, page 2015 2018, 1998.

[Pellegrini 2008] Pellegrini T., *Transcription automatique de langues peu dotées*, Thèse de doctorat, Université Paris-Sud, 2008.

[Povey 2011] Povey D., Ghoshal A., Bouliannea G., Burget L., Glembek O., Goel N., Hannemann M., Motlcek P., Qian Y., Schwar P., Silovsky J., Stemmer G. et Vesely K., The kaldi speech recognition toolkit, *IEEE Workshop on Automatic Speech Recognition and Understanding : ASRU'2011*, 2011.

[Quang 2008] Quang N. H., Nocera P., Castelli E. et Loan T. V., Reconnaissance de la parole continue á grand vocabulaire en vietnamien, une langue syllabique tonale, *Les journées d'étude de la parole : Jep'2008*, 2008.

[Quattoni 2007] Quattoni A., Wang S., Morency L.-P., Collins M. et Darrell T., Hidden conditional random fields, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 29 :1848–1853, 2007.

[Rabiner 1989] Rabiner L. et Juang B.-H., A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2) :257286, 1989.

[Rasipuram 2012] Rasipuram R. et Doss M. M., Acoustic data-driven grapheme-to-phoneme conversion using kl-hmm, *Acoustics, Speech and Signal Processing : ICASSP'2012*, pages 4841–4844, 2012.

[Rosenfeld 1994] Rosenfeld R., *Adaptive Statistical Language Modeling : A Maximum Entropy Approach*, Thèse de doctorat, Carnegie Mellon University., 1994.

[Ryan 2008] Ryan R., Rambow O., Habash N., Diab M. et Rudin C., Arabic morphological tagging, diacritization and lemmatization using lexeme models and feature ranking, *ACL (Short Papers)*, pages 117–120, 2008.

[Saadane 2015] Saadane H. et Habash N., A conventional orthography for algerian arabic, *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 2015.

[Saidane 2004] Saidane T., Zrigui M. et Ahmed M. B., La transcription orthographique-phonque de la langue arabe, *RÉCITAL' 2004*, 2004.

[Sarah 2000] Sarah L. et Sachdevref I., Code switching in tunisia : attitudinal and behavioral dimensions, *Journal of Pragmatics*, (9) :1343–61, 2000.

[Sayah 2009] Sayah M., Nagem R. et Zaghouani-Dhaouadi H., La langue arabe, histoire et controverses, *Synergies Espagne*, pages 63–78, 2009.

[Sejnowski 1987] Sejnowski T. et Rosenberg C., Parallel networks that learn to pronounce english text, *Complex Systems Publications*, pages 145–168, 1987.

[Seng 2011] Seng K., Iribe Y. et Nitta T., Letter-to-phoneme conversion based on two-stage neural network focusing on letter and phoneme contexts, *INTERSPEECH'2011, 12th Annual Conference of the International Speech Communication Association*, pages 1885–1888, 2011.

[Seng 2014] Seng K., Kouichi K., Yurie I. et Tsuneo N., Solving the phoneme conflict in grapheme-to-phoneme conversion using a two-stage neural network-based approach., *IEICE Transactions*, 97-D(4) :901–910, 2014.

[Seng 2010] Seng S., *Vers une modélisation statistique multi-niveau du langage, application aux langues peu dotées*, Thèse de doctorat, Université de Grenoble, 2010.

[Seng 2008] Seng S., Sam S., Le V.-B., Bigi B. et Besacier L., Reconnaissance automatique de la parole en langue khmère : quelles unités pour la modélisation du langage et la modélisation acoustique ?, *Les journées d'étude de la parole*, 2008.

[Sfar 2005] Sfar I., Morphologie des noms de professions : incorporation et paraphrase, *La terminologie, entre traduction et bilinguisme*, pages 156–16, 2005.

[Smaïli 1991] Smaïli K., *Conception et réalisation d'une machine á dicter á entrée vocale destinée aux grands vocabulaires : Le systéme MAUD*, Thèse de doctorat, Université de Nancy 1, 1991.

[Snoussi 2003] Snoussi F., Situation de diglossie et apprentissage de la lecture en arabe, 2003.

[Souissi 1997] Souissi E., *Étiquetage grammatical de larabe voyellé ou non*, Thèse de doctorat, Université de Paris III, 1997.

[Suontausta 2000] Suontausta J. et Häkkinen J., Decision tree based text-to-phoneme mapping for speech recognition., *Spoken Language Processing*, 2000.

[Tachbelie 2011] Tachbelie M. Y., Abate S. T. et Besacier L., Partofspeech tagging for underresourced and morphologically rich languages the case of amharic, *Conference on Human Language Technology for Development*, 2011.

[Tachbelie 2010] Tachbelie M. Y., Abate S. T. et Menzel W., Morphemebased automatic speech recognition for a morphologically rich languageamharic, *SLTU : Spoken Languages Technologies for Under-resourced languages*, 2010.

[Taravella 2011] Taravella I. E., Baude O., Maurel D., Hriba L., Dugua C. et Tellier I., Un grand corpus oral disponible : le corpus dorléans 1968-2012, *TAL*, 2011.

[Taylor 2005] Taylor P., Hidden markov models for grapheme to phoneme conversion, pages 1973–1976, ISCA, 2005.

[Tebbi 2007] Tebbi H., Transcription orthographique phonétique en vue de la synthèse de la parole rtir du texte de larabe, 2007.

[Torkkola 1993] Torkkola K., An efficient way to learn english grapheme-to-phoneme rules automatically., *IEEE International Conférence on Acoustics, Speech and Signal Processing*, 2 :199 202., 1993.

[Ueberla 1994] Ueberla J., *Analysing and Improving Statistical Language Models for Speech Recognition*, Thèse de doctorat, technishe Universitaet Muenchen, Université Joseph Fourier, Grenoble, 1994.

[Vaufreydaz 2002] Vaufreydaz D., *Modélisation statistique du langage á partir d'Internet pour la reconnaissance automatique de la parole continue*, Thèse de doctorat, Université Joseph Fourier - Grenoble, 2002.

[Vergyri 2004] Vergyri D., Kirchhoff K., Duh D. et Stolcke A., Morphologybased language modeling for arabic speech recognition, *Spoken Language Processing (ICSLP)*, pages 2245–2248, 2004.

[Vergyri 2005] Vergyri D., Kirchhoff K., Gadde V. R. R., Stolcke A. et Zheng J., Development of a conversational telephone speech recognizer for levantine arabic, *In Proceedings of INTERSPEECH'2005*, page 16131616, 2005.

[Vergyri 2008] Vergyri D., Mandal A., Wang W., Stolcke A., Zheng J., Graciarena M., Rybach D., Gollan C., Schlter R., Kirchhoff K., Faria A. et Morgan N., Development of the sri/nightingale arabic asr system, *Interspeech'2008*, page 14371440, 2008.

[Wang 2013] Wang X. et Sim K. C., Integrating conditional random fields and joint multi-gram model with syllabic features for grapheme-to-phone conversion, *NTERSPEECH 2013*, 2013.

[Wessel 2001] Wessel F., Schlter R., Macherey K. et Ney H., Confidence measures for large vocabulary continuous speech recognition, *IEEE Transactions on Speech and Audio Processing*, pages 288–298., 2001.

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

[Xu 2002] Xu P., Chelba C. et Jelinek F., A study on richer syntactic dependencies for structured language modeling, *the 40th Annual Meeting on Association for Computational Linguistics (ACL) Association for Computational Linguistics.*, pages 191–198, 2002.

[Zribi 2013] Zribi I., Khemakhem M. E. et Belguith L. H., Morphological analysis of tunisian dialect, *International Joint Conference on Natural Language Processing : IJCNLP'2013*, pages 992–996, 2013.