Outil de transcription phonétique à partir du texte Arabe

Fayçal Imedjdouben[1] – Amrane Houacine[2]

[1]Faculté d'Electronique et d'Informatique
Université des Sciences et de la Technologie Houari Boumediene Alger
f.imedjdouben@hotmail.fr

[2] Faculté d'Electronique et d'Informatique
Université des Sciences et de la Technologie Houari Boumediene Alger
ahouacine@usthb.dz

RÉSUMÉ. Nous présentons ici un outil de transcription phonétique dédié à la langue arabe. Celui-ci est basé sur une représentation selon la norme Unicode et est intégré sous l'environnement MATLAB. La transcription phonétique constitue une étape fondamentale pour la réalisation d'une synthèse à partir du texte. Le système de transcription est basé sur des règles de phonétisation et constitue un outil adapté pour la didactique et l'étude de la langue arabe.

ABSTRACT. We present a tool dedicated to the arabic phonetic transcription. It is based on the standard Unicode representation and is integrated within the MATLAB environment. The phonetic transcription is a fundamental step in building any text to speech (TTS) system. The transcription is based on phonetic rules and constitutes a suitable tool for arabic language study.

MOTS-CLÉS : Langue arabe, transcription phonétique, synthèse de la parole.

KEYWORDS: Arabic language, phonetic transcription, text to speech.

1. Introduction

La transcription phonétique consiste à représenter chaque graphème du texte d'entrée à une suite de symboles phonétiques qui seront exploités pour la production du signal acoustique.

La connaissance de la langue arabe constitue une grande partie du travail pour la réalisation d'un outil de transcription phonétique. En effet, la transcription ne peut se faire sans un travail d'analyse, de compréhension et de modélisation de la langue. Beaucoup de travaux ont été réalisés pour les langues telles que l'anglais, le français,…etc. Par contre peu de travaux sont dédiés à la transcription de la langue arabe, nous citons les travaux de Zemirli [5], Saidane [3], Al-ghamdi [1] et Ghazali [2].

Nous développons ici un outil simple de phonétisation de la langue arabe sous l'environnement MATLAB pour faciliter les travaux de synthèse en langue arabe. La norme Unicode est adoptée pour la représentation de langue arabe pour assurer la portabilité du système.

2. Etude de la langue arabe

L'alphabet de la langue arabe se compose de 28 lettres qui sont toutes des consonnes (figure 1), bien que trois d'entre elles s'emploient aussi comme des voyelles longues (ا و ي). Nous considérons pour notre part que l'alphabet arabe compte 28 consonnes et 6 voyelles (3 courtes « ـَ ـُ ـِ » et 3 longues « ـِي ـُو ـَا») et quelques réalisations vocaliques (ـْ ـّ ـً ـٍ ـٌ ـ).

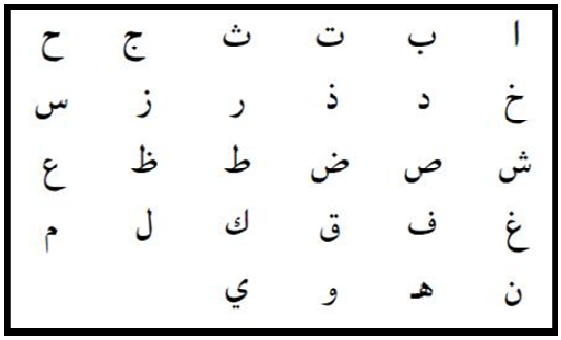

Figure 1. *L'alphabet de la langue arabe.*

La langue arabe s'écrit et se lit de droite à gauche. Les lettres arabes changent de forme de présentation selon leur position dans le mot (tableau 1).

A la fin du mot, d'une lettre non joignable	A la fin du mot, d'une lettre joignable	Au milieu du mot	Au début du mot
غ	غ	غـ	غـ

Tableau 1. *Les différentes variations de la lettre غ dans un mot.*

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au dessus ou au dessous des consonnes. L'absence des voyelles génère une certaine ambiguïté à deux niveaux :

– Sens du mot.

– Difficulté à identifier sa fonction dans la phrase.

Sept des lettres arabes s'attachent uniquement aux lettres précédentes, mais pas aux lettres suivantes. Ces lettres sont les suivantes : **ا د ذ ر ز و لا**.

Il existe deux sortes de finales : séparée (exemple : نزل) ou attachée (exemple : عمل)

Les voyelles longues ou lettres de prolongation sont les suivantes :

– Alif « ا » pour la prolongation de la consonne ayant comme voyelle courte fatha (exemple : رَا)

– Waw « و » pour la prolongation de la consonne ayant comme voyelle courte damma (exemple : رُو)

– Ya « ي » pour la prolongation de la consonne ayant comme voyelle courte kasra (exemple : ري)

Le sekun « ْ » indique que la consonne n'est pas munie de voyelle (exemple : **كُنْ**)

Le chadda « ّ » indique le redoublement de la consonne, bien qu'elle soit écrite seulement une fois (exemple : **مَدَّ**)

Le chadda « ّ » s'emploi uniquement dans les voyelles (ِ ُ َ) mais jamais avec le sekun « ْ ».

Le doublement de voyelle s'appelle tanwin : ً : an ٍ : in ٌ : un

Les lettres lunaires initiales d'un nom n'assimilent pas l'article qui les précède et par conséquent ne reçoivent pas le chadda. Ce sont : **ا ب ج ح خ ع غ ف ق ك م ه و ي**

Exemple : **اَلْقَمَرُ** ⟶ **la lettre ل est prononcée.**

Les lettres solaires initiales d'un nom assimilent l'article qui les précède et reçoivent ainsi le chadda. Ce sont : **ت ث د ذ ر ز س ش ص ض ط ظ ل ن**

Exemple : **اَلشَّمْسُ** ⟶ **la lettre ل est muette.**

Les caractères de la langue arabe n'appartiennent pas au code **ASCII**, d'où la nécessité d'utiliser un autre code qui prend en charge la langue arabe, ce code est

l'**Unicode**, ce dernier permet de coder tous les caractères utilisés par la langue arabe en mode 16 bits (tableau 2).

Unicode (Hex)	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
062		ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
063	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ					
064		ف	ق	ك	ل	م	ن	ه	و	ى	ي	ً	ٌ	ٍ	َ	ُ
065	ِ	ّ	ْ													

Tableau 2. *Standard Unicode pour les caractères arabes.*

3. Implémentation du système

Notre système de transcription phonétique est implémenté sous l'environnement **MATLAB** (langage de développement informatique particulièrement dédié aux applications scientifiques). Ce dernier fournit un environnement de programmation basé essentiellement sur le calcul matriciel, avec des fonctionnalités mathématiques et graphiques étendues.

La lecture du texte arabe se fait en mode 16 bits (en système Hexadécimal) à cause de la norme **Unicode**, l'avantage de l'utilisation de cette norme est la lecture directe du texte arabe sans avoir besoin de configurer la machine en langue arabe, ceci assure la portabilité du système de phonétisation.

Le schéma représentatif du système de phonétisation automatique est illustré dans la figure 2. La démarche que nous avons adoptée pour la réalisation du système de phonétisation automatique se décompose en deux phases de traitements linguistiques :

La première phase consiste à traiter les ponctuations, les espaces,…, de façon à ce que le texte prétraité ne comporte aucune ambiguïté pour les traitements linguistiques ultérieurs.

La deuxième phase consiste en la phonétisation du texte prétraité en utilisant 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 phonétisation correspondante aux mots sans passer par la base de règles de transcription phonétique, 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. Cette dernière utilise la norme **Unicode** pour le test des graphèmes de la langue arabe. Les règles établies (90 règles) traitent l'ensemble des

réalisations graphiques de la langue et sont au nombre de 44 graphèmes (tableau 2) pour enfin obtenir 37 phonèmes (28 consonnes, 6 voyelles, 3 réalisations vocaliques). 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. Nous obtenons ainsi le résultat phonétique du texte sans passer par une table de conversion graphème-phonème (la phonétique est incorporée dans chaque règle de la base de règle).

Nous présentons ci-dessous un exemple d'application d'une règle de transcription :
Ph : résultat phonétique.
C : caractère testé.
CD : contexte droit du caractère testé.

Règle élaborée: **[Ph] = C + CD**
Cette règle indique qu'un caractère **C**, suivi par un caractère **CD**, aura pour transcription phonétique **Ph**, soit l'exemple suivant : **[u:]** = **' ـُ ' + 'و '**
Cet exemple indique que la voyelle courte damma **' ـُ '** (représentée en Unicode par **'64F'**), suivie de la lettre waw **'و '** (représentée en Unicode par **'648'**), aura pour transcription phonétique le phonème **[u:]** selon la notation **SAMPA**.

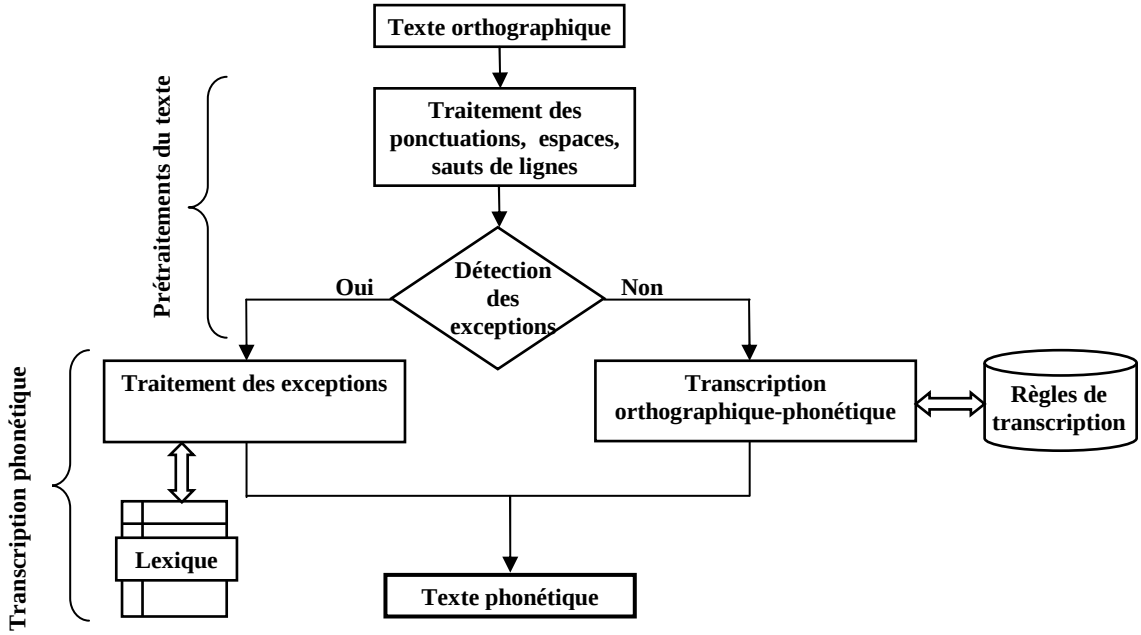

Figure 2. *Architecture du système de phonétisation automatique.*

L'interface graphique (figure 3) que nous avons développée dispose de deux possibilités pour la transcription phonétique, l'une par édition du texte directement, l'autre à partir d'un fichier. Le résultat du traitement est alors affiché en code **SAMPA** et également sous forme correspondante à la translitération en langue française. L'interface permet aussi à l'utilisateur d'obtenir des statistiques associées au texte considéré pour des études sur les aspects linguistiques et acoustiques, ainsi que des informations concernant la correspondance graphème-phonème de la langue arabe suivant la notation **SAMPA** et selon la translitération en langue française.

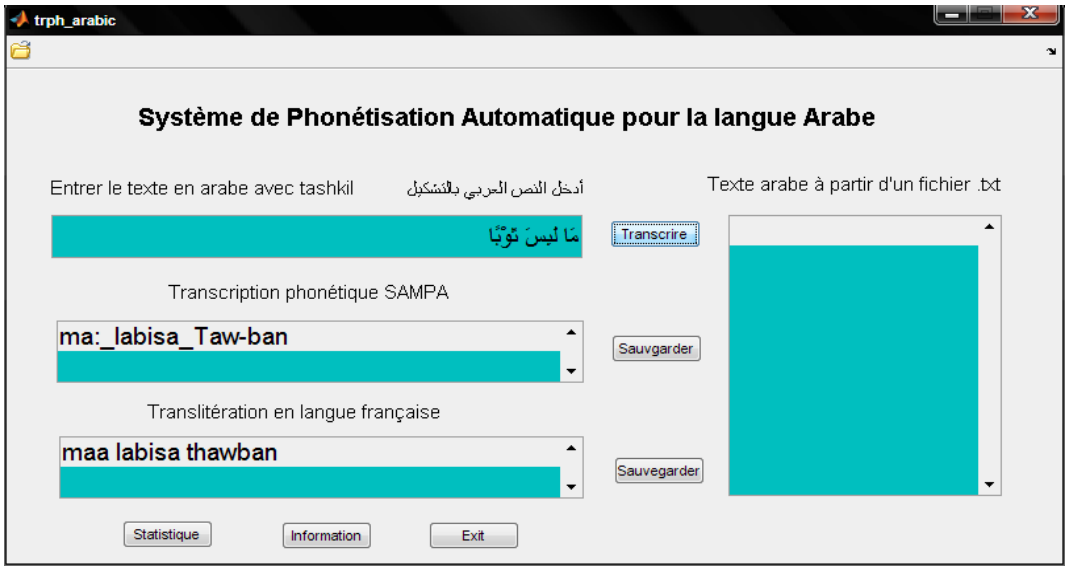

Figure 3. *Interface graphique de notre outil de phonétisation automatique.*

4. Résultats

Le système de transcription phonétique à partir du texte arabe a été testé sur une base de vingt phrases, en langue arabe, phonétiquement équilibrées [4]. Nous avons procédé à la comparaison des résultats obtenus par ce système avec ceux fournis par Saidane [4]. Les résultats de la transcription phonétique sont identiques (tableau 3). La présentation des résultats statistiques est illustrée dans les tableaux 4 et 5. Ces derniers nous fournissent des informations utiles concernant le texte à transcrire (nombre de mots, fréquence de chaque graphème du texte, type de syllabes « C/CV/CVV» constituant le texte, ainsi que leur fréquence,…).

N°	Phrase arabe	Transcription phonétique SAMPA	Translitération en langue française
1	طَفَحَ الكَيْلُ	t`afaXa_l-kaj-lu	Tafa7a lkaylu
2	أَيْنَ المُسَافِرُونَ	?aj-na_l-musa:firu:na	’ayna lmusaafiruuna
3	أَذْهَبُ بِأَمَانٍ	?aD-habu_bi?ama:nin	’adhhabu bi’amaanin
4	هَلْ لَذَعَتْهُ بِقَوْلٍ	hal_laDaH-tahu_biqaw-lin	hal ladha3tahu biqawlin
5	كُنْ هُنَا	kun_huna:	kun hunaa
6	كُنْتُ قُدْوَةً لَهُمْ	kun-tu_qud-watan_lahum	kuntu qudwatan lahum
7	لاَ لَمْ يَسْتَمْتِعْ بِثَمْرِهَا	la:_lam_jas-tam-tiH_biTam-riha:	laa lam yastamti3 bithamrihaa
8	لَمْ يَكْتُمْهُ	lam_jak-tum-hu	lam yaktumhu
9	لَوْ لاَ أَنْ مَرِضْنَا لَخَسِرُوا	law_la:_?an_marid`-na:_laxasiru:	law laa ’an mariDnaa lakhasiruu
10	قَادَ الجَيْشَ	qa:da_l-Zaj-Sa	qaada ljaycha
11	سَيُؤْذِيهِمْ زَمَانُنَا	saju?-Di:him_zama:nuna:	sayu’dhiihim zamaanunaa
12	بَعَثْتَ نَذِيرًا	baHaT-ta_naDi:ran	ba3athta nadhiiran
13	كَانَ فِي ظُلُمَاتٍ وَ لَمْ يَرْحَلْ	ka:na_fi:_D`uluma:tin_wa_lam_jar-Xal	kaana fii Zulumaatin wa lam yar7al
14	يُقَامِرُونَ بِالمَالِ	juqa:miru:na_bil-ma:li	yuqaamiruuna bilmaali
15	كَانَ صَائِمًا	ka:na_s`a:?iman	kaana Saa’iman
16	إِسْتَغْفِرْ لِذَنْبِكَ	?is-taG-fir_liDan-bika	’istaghfir lidhanbika
17	أَخَذَ إِجَازَةً	?axaDa_?iZa:zatan	’akhadha ’ijaazatan
18	لَمْ يَكُنْ شَرِسًا	lam_jakun_Sarisan	lam yakun charisan
19	لَنْ يَنْتَفِعَ	lan_jan-tafiHa	lan yantafi3a
20	مَا لَبِسَ ثَوْبًا	ma:_labisa_Taw-ban	maa labisa thawban

Tableau 3. *Résultats de la transcription phonétique pour la liste des 20 phrases.*

Syllabe C/CV/CVV	Syllabe C	Syllabe CA	Syllabe CU	Syllabe CI	Syllabe CAA	Syllabe CUU	Syllabe CII
Hamza	0	0	0	0	irréalisable	irréalisable	irréalisable
Madda	irréalisable	irréalisable	irréalisable	irréalisable	0	irréalisable	irréalisable
Alif hamza majeur	0	5	0	irréalisable	0	0	irréalisable
waw hamza	1	0	0	irréalisable	0	0	irréalisable
Alif hamza mineur	irréalisable	irréalisable	irréalisable	0	irréalisable	irréalisable	0
Ya hamza	0	0	0	1	0	0	0
Alif	0	2	0	0	0	0	0
Ba	0	2	1	6	0	0	0
Ta marbutah	irréalisable	2	0	0	irréalisable	irréalisable	irréalisable
Ta	0	5	2	2	0	0	0
Tha	1	2	0	0	0	0	0
Djim	0	1	0	0	1	0	0
7a	0	2	0	0	0	0	0
kha	0	2	0	0	0	0	0
Del	1	1	0	0	0	0	0
Dhel	1	3	0	0	0	0	2

Tableau 4. *Un extrait des statistiques des syllabes pour les 20 phrases.*

	Fréquence	Pourcentage
Mots	53	inexistant
Consonnes	207	55.2%
Voyelles courtes	100	26.6667%
Voyelles longues	22	5.8667%
Autres réalisations vocaliques	46	12.2667%
Fatha	79	21.0667%
Damma	21	5.6%
Kasra	22	5.8667%
Fathatan	6	1.6%
Dammatan	0	0,00%
kasratan	3	0.8%
Chadda	0	0,00%
Sekun	37	9.8667%
Hamza	0	0,00%
Madda	0	0,00%

Tableau 5. *Un extrait des statistiques des graphèmes pour les 20 phrases.*

5. Conclusion

Nous avons présenté ici un système opérationnel de transcription phonétique dédié à la langue arabe. Du fait de sa simplicité et de la convivialité de son interface, il constitue un outil adapté pour la didactique et la recherche sur la langue arabe.

6. Bibliographie

[1] Al-ghamdi M., Elshafei M., Al-muhtaseb H., (2002). *Arabic Text-To-Speech: Speech Units, Supported by King Abdulaziz City for Science and Technology*, 2002.

[2] Ghazali S., Habaili H., Zrigui M., *Correspondance graphème-phonème pour la synthèse de la parole arabe à partir du texte*, IRSIT Congrès dialogue homme machine, Tunis 1990.

[3] Saidane T., Zrigui M., Ben ahmed M., *La Transcription Orthographique-Phonétique de la Langue Arabe*, RÉCITAL, Fès, 19-22 avril 2004.

[4] Saidane T., Zrigui M., Ben ahmed M., *Un système de synthèse de la parole arabe par concaténation de polyphèmes : Les résultats de l'utilisation d'un lissage linéaire*, 3rd International Conférence: Sciences of Electronic, Tunis 2005.

[5] Zemirli Z., Khabet S., *Un analyseur morphosyntaxique destiné à la synthèse vocale de textes arabes voyellés*, JEP-TALN, Traitement Automatique de l'Arabe, Fès, 2004.